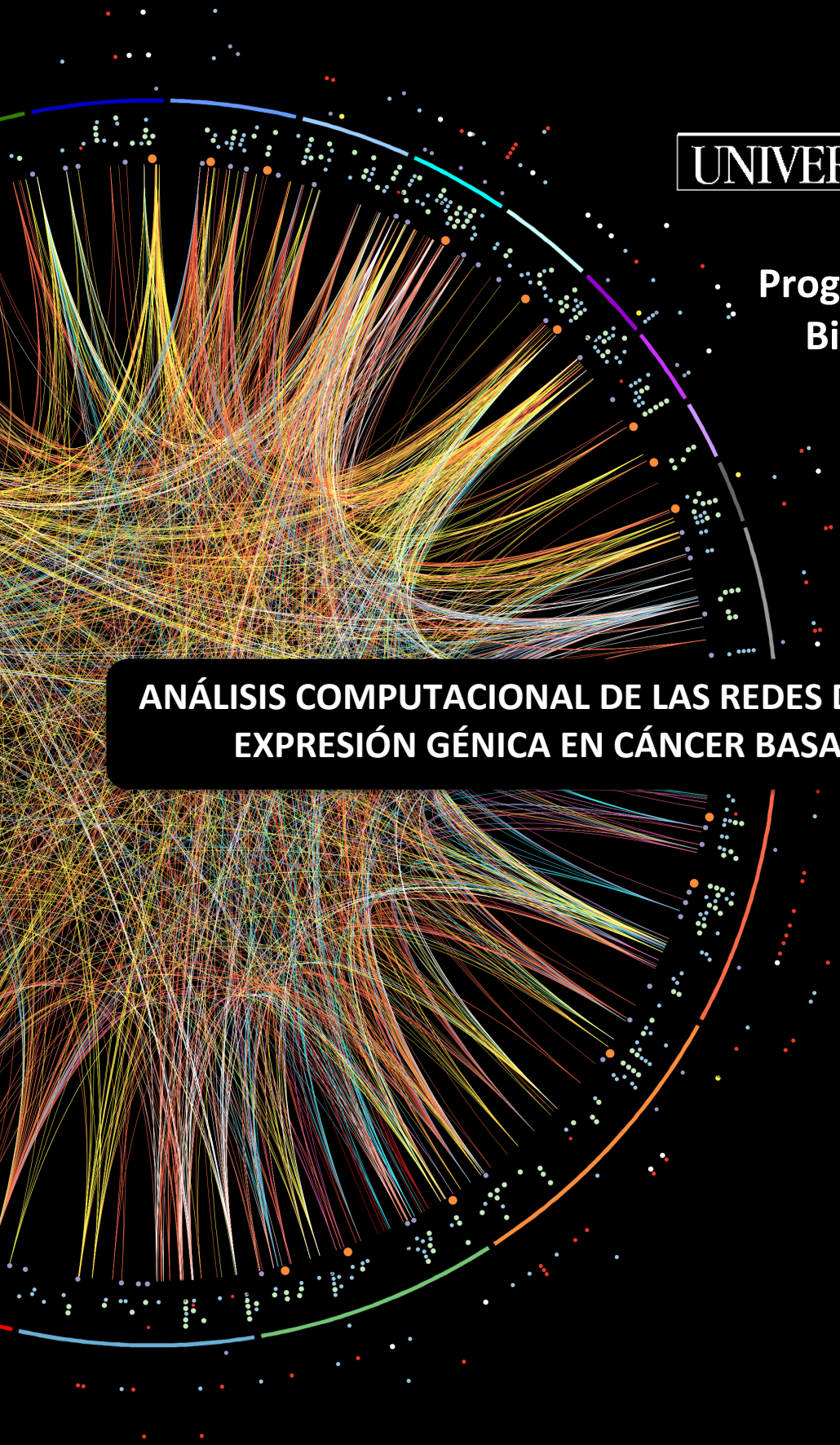


UNIVERSIDAD AUTONOMA DE MADRID

Programa de Doctorado
Biología Molecular

ANÁLISIS COMPUTACIONAL DE LAS REDES DE REGULACIÓN DE LA EXPRESIÓN GÉNICA EN CÁNCER BASADAS EN MIARNS

Eduardo Andrés León

Madrid, 2017

Departamento de Biología Molecular

Facultad de Ciencias

Universidad Autónoma de Madrid

ANÁLISIS COMPUTACIONAL DE LAS REDES DE REGULACIÓN DE LA EXPRESIÓN GÉNICA EN CÁNCER BASADAS EN MIARNS.

Tesis doctoral que presenta para optar al grado de doctor en Ciencias por la Universidad Autónoma de Madrid, el licenciado en Ciencias Biológicas

Eduardo Andrés León

Co-Directores de Tesis:

Dra. Ana María Rojas Mendoza

Prof. Alfonso Valencia Herrera

AGRADECIMIENTOS

Siempre he escuchado que redactar los agradecimientos de una tesis es la parte más fácil y bonita. Tal es así, que me he imaginado muchas veces dedicando estas palabras a los que me han ayudado en este largo y tortuoso camino. Sin embargo, llegado el momento no sé ni por donde, ni por quien empezar. En primer lugar, mis más sinceros agradecimientos a los directores de esta tesis, ya que sin ellos este estudio no hubiera visto la luz. En especial a Alfonso Valencia, tengo muchísimo que agradecerte, tu fuiste el primero en darme una oportunidad, y en confiar en mí para abrirme paso en este mundo de la Bioinformática. Ana Rojas, a ti te doy un millón de gracias, ya que con tu ímpetu y tus inagotables ganas, al final hemos logrado hacer un gran trabajo juntos. En definitiva, sin ti esta tesis no existiría.

Por supuesto, no puedo olvidarme de mis compañeros de trabajo, vosotros estáis de una forma u otra presentes en esta tesis, porque he tenido la suerte de aprender mucho de vosotros. Recuerdo al principio cuando Christian, Fede, Juan Carlos, José María y José Manuel me explicaban como funcionaba Perl y las máquinas Linux. Más tarde, en el CNIO, tuve la suerte de trabajar de cerca con Gonzalo, Pisano, y entre todos, creamos miRGate, la semilla que dio lugar a esta tesis. También dar las gracias a muchas más personas con las que he colaborado en mayor o menor medida, como Osvaldo, Willy, Fátima, Andrés, Dani, Enrique, el pelirrojo, Izaskun, Raquel, César, Michael, Paolo …y muchos más, que seguro que ahora mismo, no recuerdo. ¡Gracias de corazón!.

Un párrafo entero se merece mi compañero y sobretodo amigo, Ángel. Mucho de lo que sé ahora sobre hashes, referencias y un montón de cosas más, te las debo a ti. Gran parte del código creado en esta tesis, lo aprendí de ti. Además, muchas gracias en el ámbito personal por esas resacas, fiestas y momentos inolvidables en mi vida. Suerte en Roatán.

También quisiera agradecer a mis amigos vuestro apoyo siempre que ha hecho falta, Rubén, Laura y Vanesa, os agradezco vuestra inestimable ayuda en las etapas más tristes de mi vida. A los walking, por hacerme olvidar todo, sea lo que fuere. A Elena y Rocío, mis andaluzas favoritas, con vosotras el trabajo y la amistad pasan a otro nivel, gracias por esas sevillanas regadas de rebujito. Me gustaría agradecer a Reyes su insistencia para enviar un CV, que pensé que nunca llegaría a ninguna parte, gracias a ti, con ese CV conseguí mi primer trabajo, y luego toda una carrera profesional.

Por último, quisiera agradecer a mi paciente familia su apoyo incondicional durante tantos años, sobretodo en estos últimos, de trabajo intenso. Gracias por entender mis constantes ausencias en momentos familiares importantes. Elena, gracias hermana por encargarte de tantos temas que yo no he podido, y gracias por regalarme dos sobrinos tan maravillosos. Papa, Mamá, gracias por los esfuerzos que habéis hecho siempre, por enseñarme a ser

quien soy ahora. Hortensia, a ti infinitas gracias por creer en mí, por insistir y animarme sin descanso a alcanzar el grado de doctor.
Antes de finalizar, quisiera dar las gracias a la persona más importante en mi vida. Amor mío, gracias por aguantar lo inaguantable, por no desesperar, por consolarme y ayudarme tantas veces. Los dos sabemos que sin tu ayuda y paciencia, esta tesis ni hubiera empezado. Raquel un millón de gracias. Te amo.

GRACIAS A TODOS DE CORAZÓN!!

RESUMEN / SUMMARY

El cáncer, con más de 200 tipos tumorales distintos conocidos, representa la tercera patología con mayor índice de mortalidad en el mundo. Éste se caracteriza por una elevada tasa de crecimiento, la evasión de barreras anti-proliferativas, su capacidad de replicación descontrolada y la invasión a tejidos cercanos. Vinculado a estas características, encontramos distintas rutas de regulación como son: el ciclo celular, la ruta asociada al daño en el ADN, la elongación de telómeros, la replicación del DNA, la senescencia y la muerte celular (ya sea por apoptosis o necrosis), etc.

Aunque el origen de un tumor es un proceso en el que intervienen numerosos factores, muchos aún desconocidos, se ha observado que tanto las modificaciones genéticas como las epigenéticas en el ADN, confieren una mayor susceptibilidad a desarrollar esta enfermedad. Tal es así que alteraciones como las mutaciones, la variación en el número de copias, la metilación, la accesibilidad de cromatina o la regulación por microARNs, son factores íntimamente relacionados con el desarrollo y el progreso tumoral. Una extensa investigación sobre los mecanismos moleculares de la tumorogénesis, ha llevado a la caracterización de oncogenes y supresores de tumores que son elementos clave en el crecimiento y la progresión del cáncer, así como la de otros elementos importantes como son los microARNs. Estos genes y miARNs aparecen constitutivamente desregulados en el cáncer.

El objetivo principal de esta tesis es desarrollar métodos computacionales que permitan identificar nuevas interacciones entre miARNs y genes desregulados, conservadas en las rutas asociadas al cáncer. Para ello hemos desarrollado una herramienta que permite predecir sitios de unión entre miARNs y ARNm con una fiabilidad mayor de la disponible hasta esa fecha. Posteriormente se desarrolló una segunda utilidad que permite procesar de forma eficiente y automática muestras de pacientes para, de esta forma, obtener aquellos genes y miARNs desregulados. El uso conjunto de ambas herramientas nos permitió llevar a cabo un análisis de 18605 muestras de transcriptoma, procedentes de 15 de los tipos de tumores más comunes, teniendo en cuenta el efecto de la metilación y de la alteración en el número copia génicas y su influencia en la desregulación de los genes.

A partir de este estudio de transcriptoma global, hemos recuperado interacciones conocidas por su papel en el desarrollo tumoral, así como nuevas asociaciones formadas principalmente por oncogenes y miARNs supresores de tumores que en algunos casos además, están asociadas con la supervivencia de los pacientes. Creemos que los resultados presentados en esta tesis podrían facilitar el diseño de fármacos no cito-tóxicos orientados a paliar la desregulación tanto de genes como de microARNs cuyos cambios de expresión están vinculados con la tumorogénesis o la baja supervivencia.

Cancer, with more than 200 different tumour types known, represents the third most common pathology with the highest mortality rate worldwide. It is characterized by an unrestrained cell proliferation, the avoidance of growth suppressors, resistance to cell death, activation of metastasis and initiation of replicative immortality. Several regulatory pathways emerge related to these hallmarks, such as: cell cycle, DNA damage response, elongation of telomeres, DNA replication, senescence and cell death (either apoptosis or necrosis), etc.

Although the origin of a tumour is a process involving several different factors, many unknown, it has been observed that genetic and epigenetic modifications affecting DNA, increases the susceptibility to develop cancer. Thus, variations such as mutations, alteration in gene copy members, methylation, accessibility of chromatin or microRNA regulation, are factors closely related to tumour development and progression. Extensive research into the underlying molecular mechanisms of tumorogenesis has led to the characterization of oncogenes and tumour suppressors that are key elements in the growth and progression of cancer, as well as other important elements such as microRNAs. These genes and miRNAs appear constitutively deregulated in cancer.

The main objective of this thesis is to develop novel computational methods to identify new miRNA-gene interactions conserved in cancer-associated pathways. To this end, we have developed a tool able to predict highly reliable miRNAs-target sites not covered by others resources. Subsequently, a second utility was developed to efficiently process patient samples, in order to obtain differentially expressed cancer genes and miRNAs. The combined use of both tools allowed us to carry out an analysis of 18,605 raw transcriptome samples from 15 of the most common tumours types, accounting for methylation and the alteration in gene copy number effects and their influence on gene deregulation.

From this global transcriptome analysis, we have recovered interactions known for their role in tumour development, as well as new associations mainly constituted by oncogenes and tumour suppressor miRNAs that, in some cases, are associated with patient survival too. We believe that the results presented in this thesis could provide relevant information regarding the signature's stability in cancer-related pathways and patient survival, and such information is likely to be useful to define novel therapeutic strategies.

ÍNDICE

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ABREVIATURAS

ANOVA	Análisis de varianza.
API	Interfaz de Acceso Programático.
ARNm	ARN mensajero.
AUC	Área bajo la curva, del inglés Area Under the Curve.
BAM	Fichero binario que contiene datos de secuencias alineadas frente al genoma de referencia.
CDK	Ciclinas dependientes de quinasas.
CDKI	Inhibidores de ciclinas dependientes de quinasas.
CNAs	Alteración en el número de copias génicas.
CPM	Número de secuencias por cada millón de lecturas alineadas, del inglés Counts Per Million.
CWS	Valor pesado y ponderado, del ingles consensus weighted score.
CpG	Citosina seguida de una guanina.
DDR	Respuesta al daño en el ADN.
DE	Expresión diferencial.
DSB	Rotura de ambas hebras del DNA.
FDR	Falso ratio de descubrimiento.
GEO	Repositorio de muestras experimentales. Del inglés Gene Expression Omnibus.
GRC	Consorcio del genoma de referencia.
HGNC	Comité de nomenclatura de genes. HUGO Gene Nomenclature Commitee, en inglés.
HUGO	Organización del genoma humano, del inglés Human Genome Organisation.
KEGG	Base de datos de rutas metabólicas anotada manualmente.
LogFC	Cambio de expresión entre dos grupos de muestras en logaritmo en base dos.
miARN	micro ARN.
NGS	Secuenciación de nueva generación. Del inglés, Next Generation Sequencing.
OD	Razón de probabilidades, del inglés odd ratio.
pSILAC	Ensayo basado en el marcaje de isótopos estables en aminoácidos en un cultivo celular.

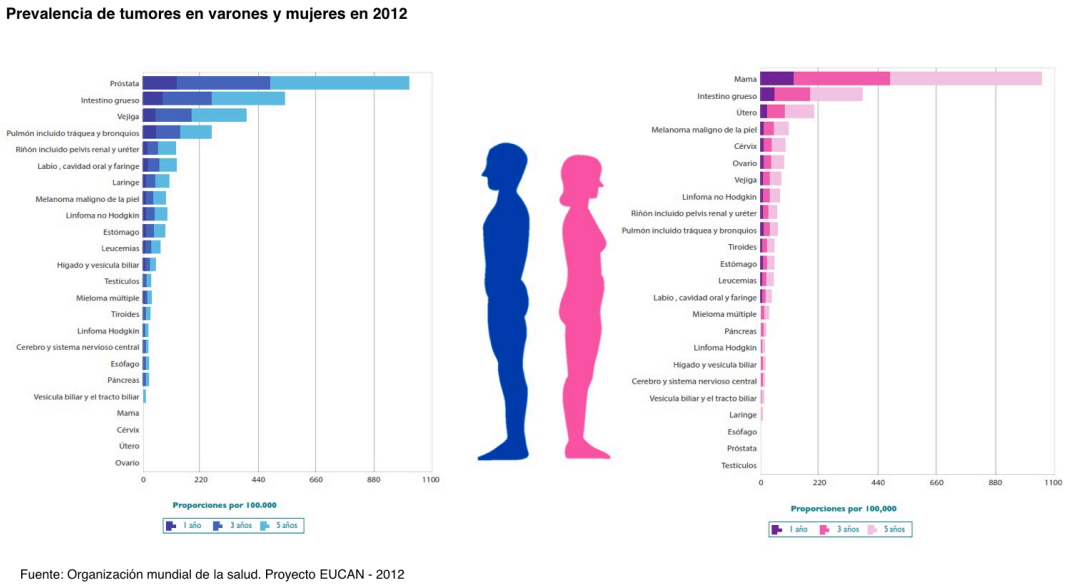
REST	Transferencia del estado representacional, Representational state transfer, en inglés.
RNAi	Ruta de ribo-interferencia.
RNS	Especie reactiva de nitrógeno (del inglés Reactive Nitrogen specie).
ROC	Característica Operativa del Receptor, del inglés receiver operating characteristic.
ROS	Especie reactiva de oxígeno (del inglés Reactive Oxygen specie).
RPKM	Secuencias por cada mil nucleótidos de transcrito por cada millón de secuencias alineadas, del inglés Reads Per Kilobase of transcript per Million mapped reads.
SRA	Repositorio de muestras experimentales procedentes de tecinas de secuenciación de nueva generación. Del ingles Sequence Read Archive.
TCGA	Altas Genómico del Cáncer.
TLS	Síntesis de DNA de translesión.
TMM	Algoritmo de normalización de RNASeq basado en una distribución binomial negativa, del ingles trimmed mean of M-values.
UTR	Región no traducida.
XML	Lenguaje de marcas extensible, del inglés eXtensible Markup Language.

1. INTRODUCCIÓN

1.1 Cáncer.

Cáncer es un término genérico que engloba a un conjunto de enfermedades que pueden desarrollarse en cualquier parte del organismo, también denominadas neoplasias malignas o tumores malignos. Actualmente se han descrito más de 200 tipos de tumores distintos.

En un estudio llevado a cabo en 2012 (Ferlay, Steliarova-Foucher et al. 2013) por la agencia internacional del cáncer, departamento dependiente de la Organización Mundial de la Salud, el cáncer fue considerado como una de las principales causas de morbilidad y mortalidad a nivel mundial. Sólo en 2012 se diagnosticaron 14 millones de nuevos casos y 8.2 millones de muertes, se debieron a este conjunto de enfermedades. En ese mismo año, los tipos de tumor diagnosticados con mayor frecuencia en España fueron los de próstata, colon, recto, vejiga y pulmón (Figura 1) en el caso del hombre, mientras que en mujeres fueron los de mama, colon, recto, cuello uterino y melanoma (Figura 1). Además, de ese estudio se desprende que en las próximas dos décadas, el número de nuevos tumores diagnosticados se eleve hasta cerca de los 22 millones.



Fuente: Organización mundial de la salud. Proyecto EUCAN - 2012

Figura 1. Prevalencia de los distintos tipos de tumores en varones y mujeres en el año 2012 según la agencia internacional el cáncer (Organización Mundial de la Salud) en España. Adaptado de http://eco.iarc.fr/eucan/.

1.2. Agentes genéticos y epigenéticos causales del cáncer.

La células que constituyen nuestros órganos, a lo largo de su periodo funcional, reciben diferentes señales que determinarán la división, la diferenciación o la muerte celular. Una de las propiedades fundamentales de las células tumorales es su capacidad de evadir estas

señales, desencadenando así un crecimiento descontrolado y la proliferación celular. En el caso de que esta proliferación celular continuara, podría propagarse a tejidos circundantes y tendría lugar el proceso denominado metástasis, que es el causante del 90% de las muertes asociadas al cáncer.

El origen de un tumor es un proceso muy complejo en el que intervienen una elevada cantidad de factores diversos, muchos de ellos aún desconocidos. Asimismo, es cierto que se conocen agentes químicos con capacidad carcinogénica, entre estos, podemos destacar aquellos generados por nuestro propio organismo, como son los procedentes del metabolismo celular o la microbiota intestinal (Voulgaridou, Anestopoulos et al. 2011). A su vez, la activación de la células del sistema inmune, tales como monocitos y macrófagos (Fialkow, Wang et al. 2007) originan radicales potencialmente tóxicos. Por otra parte, dentro del grupo de los agentes genotóxicos externos, destacamos especialmente la radiación ultravioleta (Kielbassa, Roza et al. 1997), la radiación ionizante (Dipple 1995, Cadet, Ravanat et al. 2012), el tabaco (Sasco, Secretan et al. 2004), las infecciones víricas (Martin-Perez, Vargiu et al. 2012) y otros agentes presentes en la comida, el agua, aire o los fármacos (Roos and Kaina 2013). Debido a que las células epidérmicas del tracto respiratorio y del aparato digestivo son las más expuestas a estos elementos carcinogénicos, no sorprende que el 90 % de los tumores se originen en células epiteliales, conocidos como carcinomas. En resumen, todos estos elementos atacan al ADN celular, provocando una extensa variedad de lesiones (Dipple 1995), entre las que destacan las mutaciones (cambios en la secuencia del ADN), los daños cromosómicos responsables de una expresión génica aberrante, la producción de variantes génicas anormales o la alteración en los patrones de metilación de promotores que, entre otros, eventualmente ocasionan la transformación oncogénica y la progresión tumoral.

Además, es importante destacar que entre un 5-10% de los tumores diagnosticados, tienen un origen genético debido a la herencia paterna o materna de genes que presentan alteraciones en su secuencia, como son las mutaciones. Éstas son cambios que afectan a los nucleótidos que definen a un gen y cuya variación puede provocar la síntesis de proteínas no funcionales y desembocar así en un fenotipo oncogénico.

En los tejido tumorales además de mutaciones, se han observado cambios de secuencia pero, que afectan a grandes porciones del genoma y que propician alteraciones en el número de copias génicas (del inglés CNAs, *Copy Number Alterations)*. Estas grandes variaciones, favorecen en unos casos la ganancia de copias de genes y por lo tanto beneficiando el aumento de su expresión, como el asociado a genes formadores de tumores

(Wang, Lim et al. 2013), la perdida de copias (asociado a la represión de genes supresores de tumores (Krepischi, Maschietto et al. 2016)), e incluso la perdida de heterocigosidad (Chen, Zhang et al. 2015).

Junto a estas alteraciones genéticas presentes en los tumores, el sucesivo acúmulo de modificaciones epigenéticas, confiere de igual manera una mayor susceptibilidad al desarrollo tumoral. Dentro de la epigenética (cambios heredables en la expresión génica no asociada a la modificación de la secuencia del ADN) podemos distinguir dos tipos principales de eventos asociados en cáncer. Estos son, los cambios conformacionales de la estructura de la cromatina asociados a la modificación química de las histonas mediante metilaciones o acetilaciones y, la metilación del ADN. Éste último se caracteriza por la metilación en el genoma celular de las citosinas que preceden a una guanina, los llamados dinucleótidos CpGs. Estos dinucleótidos se caracterizan por concentrarse formando lo que se conoce como islas CpG (Generalmente definidas como regiones de ADN de 1000 kilobases con un contenido de GC superior al 50%). El 60% de estas islas se encuentran en las regiones promotoras de los genes (Vu, Li et al. 2000) y son capaces de regular la expresión del gen asociado (Herman and Baylin 2003, Weber, Hellmann et al. 2007) según el estado de metilación que presenten las citosinas que forman las islas CpGs. Por ejemplo el supresor tumoral *PTEN*, presenta bajos niveles de expresión tanto en los tumores de cerebro como en los de tiroides, dado que su región promotora suele aparecer fuertemente metilada en comparación con el tejido sano (Virani, Colacino et al. 2012). Sin embargo, el oncogen *INSL4*, también en cáncer tiroides, presenta una hipo-metilación estadísticamente significativa comparada con el tejido sano (Rodriguez-Rodero, Fernandez et al. 2013).

Dentro de los factores epigenéticos capaces de modificar la expresión génica, algunos investigadores engloban además, la regulación llevada a cabo por los microARNs. Estos ejercen su función en forma de pequeños ARNs mono-catenarios modificando la expresión de los genes mediante su unión a regiones con secuencias complementarias. El mecanismo de acción más conocido, se basa en la represión de la expresión génica, mediante la unión a regiones en el 3'-UTR (del inglés *Untranslated Region*), aunque se han descubierto microARNs capaces de unirse a secuencias promotoras en 5'-UTR e inducir la expresión de genes (Place, Li et al. 2008).

Los microARNs son objeto de un profundo estudio en esta tesis doctoral debido a su implicación en cáncer y a la posibilidad que ofrecen para el desarrollo de terapias no cito-tóxicas.

1.3. Señas de identidad del cáncer.

El cáncer como conjunto diverso de tipos tumorales, es una enfermedad muy compleja tanto a nivel celular como molecular. Sin embargo, D. Hanahan y R. Weinberg en el año 2000 resumieron a seis (Figura 2), las principales capacidades biológicas, que aunque distintas y complementarias, permiten el crecimiento tumoral y la invasión a órganos cercanos desencadenando el proceso de metástasis (Hanahan and Weinberg 2000, Hanahan and Weinberg 2011). Posteriormente, en 2011 (Hanahan and Weinberg 2011), ambos autores sugirieron la inclusión de la reprogramación del metabolismo energético y la evasión del sistema inmune, como dos factores relacionados tanto con el crecimiento como con el desarrollo tumoral. A continuación se expone la información fundamental de estas características.

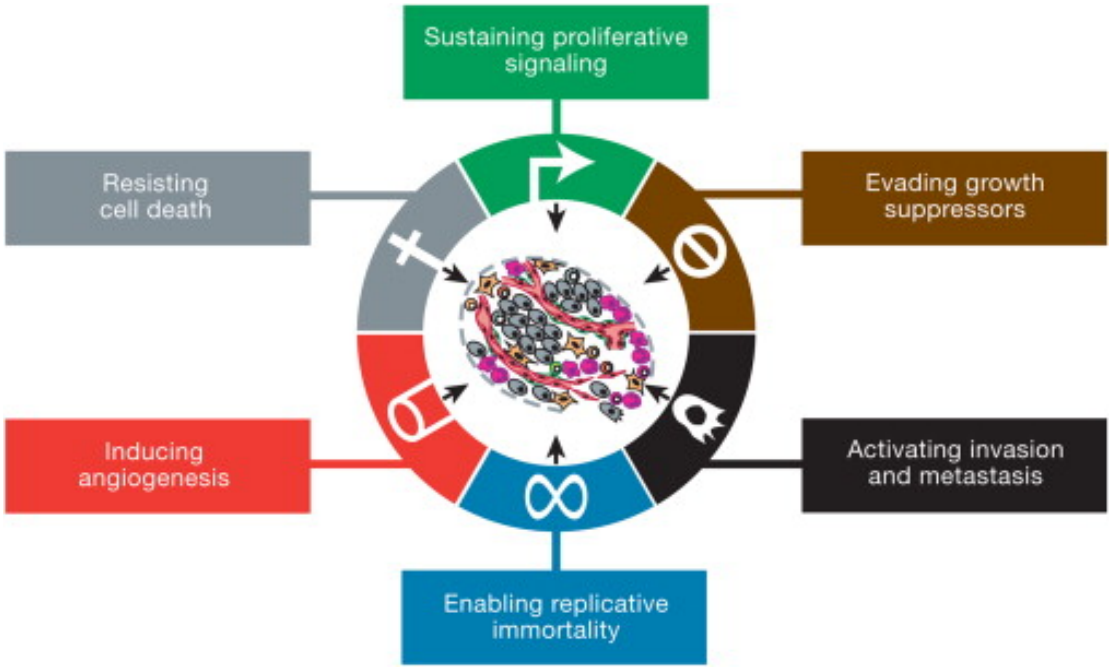


Figura 2. Características principales del origen tumoral. Adaptado de Hanahan et al. Cell (2011).

1.3.1. Proliferación celular.

Una particularidad fundamental de las células tumorales, es su capacidad de sostener una proliferación crónica. Las células que conforman los tejidos sanos, controlan de forma precisa la producción y liberación de señales promotoras del crecimiento, que indican la entrada y la progresión a través del ciclo celular y la división. Por el contrario, las células tumorales incrementan la expresión de factores de crecimiento, desencadenando la activación de vías de señalización intracelulares que regulan la progresión a través de la división, así como el crecimiento celular. De esta forma, se conoce que las células

tumorales son capaces de producir ligandos asociados a factores de crecimiento que favorecen la estimulación proliferativa autocrína (Cheng, Fan et al. 2008). Alternativamente, además pueden enviar señales de estímulo a través del estroma a células normales, con el objetivo de obtener factores de crecimiento exógenos adicionales(Bhowmick, Neilson et al. 2004).

1.3.2. Evasión de las señales de inhibición del crecimiento.

Las células tumorales, con el fin de estimular el crecimiento y la progresión tumoral, además de sobre-expresar la señales de crecimiento recién comentadas, precisan de forma adicional, eludir las vías de regulación dependientes de los denominados genes supresores de tumores, los cuales impiden la proliferación celular exacerbada. Entre estos genes, destaca la proteína de retinoblastoma (RB) y el factor de transcripción p53. RB, interacciona con numerosas proteínas, pero su unión a los miembros de la familia de factores de transcripción E2F, parece ser su papel principal en el control de la replicación celular (Dyson 1998, Yamasaki 1998). Por otra parte, el factor de transcripción p53 es inducido en respuesta al daño del ADN, la hipoxia y la activación de oncogenes. P53 también regula un programa de expresión génica que conduce a la detención del ciclo celular o a la apoptosis (Levine 1989, Giaccia and Kastan 1998). Entre los genes inducidos por p53 es importante resaltar a *CDKI p21* y genes que codifican proteínas con funciones pro-apoptóticas. Interesantemente, *TP53* aparece mutado en más del 50% de los distintos tipos de cánceres humanos, y mutaciones en genes que afectan la función de *TP53* se producen en muchos, si no todos, los tumores con un *TP53* normal.

1.3.3. Inmortalidad.

Las células tumorales se caracterizan por su capacidad de replicación casi indefinida, llegando a generar durante este proceso tumores macroscópicos. Por el contrario, las células sanas se dividen un número limitado de veces hasta entrar en un estadío celular viable no proliferativo, denominado senescencia. En el caso de escapar a este proceso, reciben señales moleculares que conllevan a su muerte celular. Las células tumorales, sin embargo, son capaces de evadir la activación de estas barreras naturales frente a la proliferación celular. En este contexto, diversos investigadores defienden el papel de los telómeros como los principales responsables de este proceso. Estos estudios se basan en los elevados niveles de expresión de la telomerasa (proteína encargada de la síntesis de los telómeros) presentes en las células tumorales, a diferencia de la nula expresión de esta enzima en células normales. También se ha demostrado que la presencia de actividad

telomerasa, correlaciona con la resistencia tanto a la entrada en senescencia como a apoptosis. De esta forma, la inhibición de la actividad telomerasa, desencadena el acortamiento de los telómeros, y por consiguiente, la activación de las barreras anti-proliferativas (Blasco 2005).

1.3.4. Resistencia a la muerte celular programada.

El proceso de muerte celular programada representa una barrera natural frente al desarrollo sin control y al crecimiento del cáncer. Aunque las condiciones celulares necesarias para desencadenar la apoptosis no han sido totalmente elucidadas, diversos autores han demostrado la existencia de varios sensores con un papel crucial en la evasión de la apoptosis y por consiguiente, el posterior desarrollo tumoral. Entre ellos, el más conocido es p53, el cual induce apoptosis mediante la sobre-expresión de *Noxa* y *Puma*, de forma que es común la perdida total o la disminución de los niveles de expresión de p53 en tumores. Adicionalmente, también se ha esclarecido que el incremento de los niveles de expresión de genes anti-apoptóticos o la inhibición de la expresión de factores pro-apoptóticos de la familia Bcl-2, pueden impedir la activación de la muerte celular. En concreto, *Bcl-2* junto con sus parientes más cercanos (*Bcl-xL, Bcl-w, MCL-1, A1*) reprimen la activación de la apoptosis a través de la unión a los dominios de represión BH3 de dos proteínas pro-apoptóticas (Bax y Bak) (Adams and Cory 2007).

1.3.5. Angiogénesis.

Al igual que los tejidos sanos, los tumores requieren del aporte de nutrientes y oxígeno, así como la eliminación de desechos metabólicos. Durante la embriogénesis, la formación de vasculares sanguíneos favorece este intercambio de sustancias con el objetivo de favorecer el desarrollo. Esto fomenta el crecimiento de nuevas células endoteliales y su ensamblaje en forma de tubos (proceso denominado vasculogénesis), además del desarrollo de nuevos vasos a partir de los existentes, denominado angiogénesis. En el organismo adulto, este proceso permanece inactivo y sólo se impulsa de forma transitoria en respuesta a heridas o a ciclos reproductivos. Sin embargo, se ha descrito un proceso de neo-vascularización relacionado con el crecimiento de tumores, dependiente de factores de crecimiento como el *bFGF* y *VEGF*. De esta forma, la angiogénesis permite proveer los nutrientes necesarios para el desarrollo tumoral, así como para la diseminación del cáncer. Esto es debido a que las células cancerosas pueden desprenderse de un tumor sólido determinado, entrar en un vaso sanguíneo o linfático y trasladarse a un sitio distante, donde pueden implantarse y comenzar el crecimiento de un tumor secundario o metástasis. Asociado a este proceso

encontramos el aumento de la expresión de factores de crecimiento endotelial como *VEFG-A* o la represión de inhibidores como *TSP-1* (Baeriswyl and Christofori 2009).

1.3.6. Activación de la invasión tumoral y la metástasis.

El programa regulador que proporciona la competencia de invadir tejidos, depende fundamentalmente del tipo celular originario del tumor. En el caso de los carcinomas, la transición conocida como epitelio-mesénquima es el proceso por el cual las células epiteliales experimentan determinadas transformaciones que les confieren la capacidad de invadir, diseminarse y resistir a la apoptosis (Barrallo-Gimeno and Nieto 2005). Entre estas modificaciones, destacamos la alteración de los niveles de expresión de la E-caderina, molécula encargada de la adhesión molecular célula-célula. De tal forma, niveles bajos de E-caderina son propios de tumores altamente invasivos y metastásicos, por el contrario, su sobre-expresión impide la diseminación tumoral.

1.4. Rutas génicas asociadas a las señas de identidad del cáncer.

Estas características presentes en todos los tipos de tumores, aunque generales, nos permiten profundizar en los mecanismos moleculares implicados en la tumorogénesis. De esta forma y dependiente de ellas, podemos definir varias rutas de regulación, cuyos proteínas implicadas, están íntimamente relacionadas con la formación y progresión tumoral. Esto es debido a que los genes que codifican estas proteínas sufren modificaciones tanto genéticas como epigenéticas que alteran su correcta función. Entre ellos a continuación resaltaremos el ciclo celular, la respuesta al daño en el ADN, la replicación del ADN y la elongación de los telómeros, dado que las alteraciones sufridas sobre los genes implicados en ellas, favorecen la progresión tumoral. De esta forma, estas rutas están enriquecidas en oncogenes. A su vez, también puede desencadenarse la inhibición de la expresión funcional de genes supresores de tumores, que tal y como se ha comentado, están implicados en la parada del ciclo celular, la senescencia y la muerte celular (principalmente apoptosis o incluso necrosis).

1.4.1. Ruta de la respuesta al daño en el ADN.

El origen del cáncer como se ha comentado, esta relacionado con la acumulación sucesiva de alteraciones (Figura 3) que desencadenan un desequilibrio, que favorece la transformación oncogénica y la progresión tumoral (Miller and Miller 1981). Para controlar la inestabilidad genómica, las células activan distintas rutas de respuesta ante tales daños en el ADN. Concretamente, estas rutas son capaces de reparar, eliminar o incluso en

ocasiones, tolerar estas lesiones genómicas (Arcas, Fernandez-Capetillo et al. 2014). Por el contrario, la no reparación de los daños y la consecuente acumulación de alteraciones, favorece la eliminación de la célula mediante el proceso de apoptosis o de muerte por necrosis (Hoeijmakers 2001).

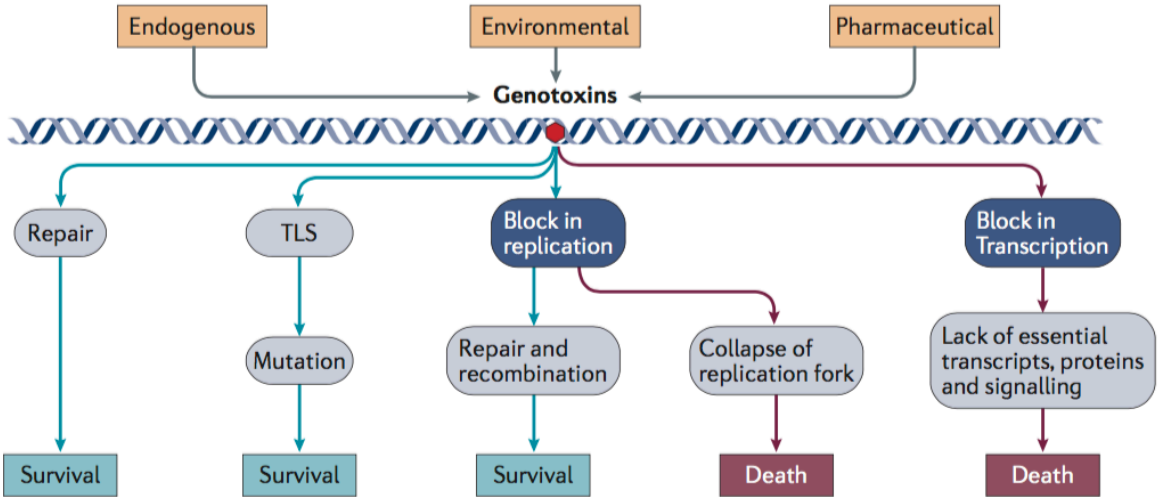


Figura 3. Ejemplo de distintos agentes genotóxicos que producen múltiples tipos de daños en el DNA. Dependiendo de la magnitud de los daños, éstos pueden ser tolerados, reparados o por el contrario letales a nivel celular. Adaptado de Wynand Roos et al. Nature Reviews cancer (2016).

Los distintos niveles de daño parecen generar diferentes respuestas en las células. Este paradigma establece que bajos niveles de daño en el ADN desencadenan una respuesta de reparación y supervivencia, mientras que niveles altos, bloquean el ciclo celular con el propósito de reparar tales daños. En el caso en que estos daños no puedan ser reparados, se desencadena una cascada de señalización que conlleva a la muerte celular.

Numerosos genes de diversas rutas participan en este paso, por ejemplo *TP53* contribuye a la expresión diferencial de genes pro-supervivencia o pro-apoptosis (Tian, Liu et al. 2012). Por otra parte, tanto *ATM* como *ATR* promueven la activación de la muerte celular en respuesta a elevados niveles de daño en el ADN induciendo a *CASP2, E2F1, P73 o CHCK1* (Xu and Baltimore 1996). Por el contrario, aquellos daños producidos por especies reactivas de oxígeno (ROS) o de nitrógeno (RNS) activan a *PARP-1* (poli-ADP-ribosa polimerasa I), que inicia la muerte celular por necrosis (Los, Mozoluk et al. 2002).

1.4.2. Ciclo celular.

El ciclo celular es el proceso en el que una célula se divide en dos células hijas idénticas genéticamente a la célula primigenia y se compone de cuatro fases secuenciales denominadas: i) fase G1, etapa funcional de la célula que finaliza al activarse señales de

entrada en la ii) fase S, en la cual tiene lugar la replicación del ADN. A continuación la iii) fase G2 prepara a la célula para su entrada en la última iv) fase, llamada M o mitosis, donde la célula progenitora se divide finalmente en dos células. La progresión a través de las distintas etapas del ciclo celular está controlada por la expresión de un conjunto de proteínas llamadas Ciclinas Dependientes de Quinasas (CDK) además, de las proteínas que las regulan, como son las proteínas Inhibidoras de las Ciclinas Dependientes de Quinasas (CDKI). De tal forma que la unión de la ciclina-D con CDK4 y CDK6, junto a la interacción de la ciclina-E con CDK2 desencadenan el paso de la fase G1 a la fase S. Igualmente, la fase S se inicia gracias a la unión de la ciclina-A con CDK2 y por último, la asociación de la ciclina-B junto a CDK1, regula la progresión en la fase G2, así como la entrada en mitosis (Williams and Stoeber 2012).

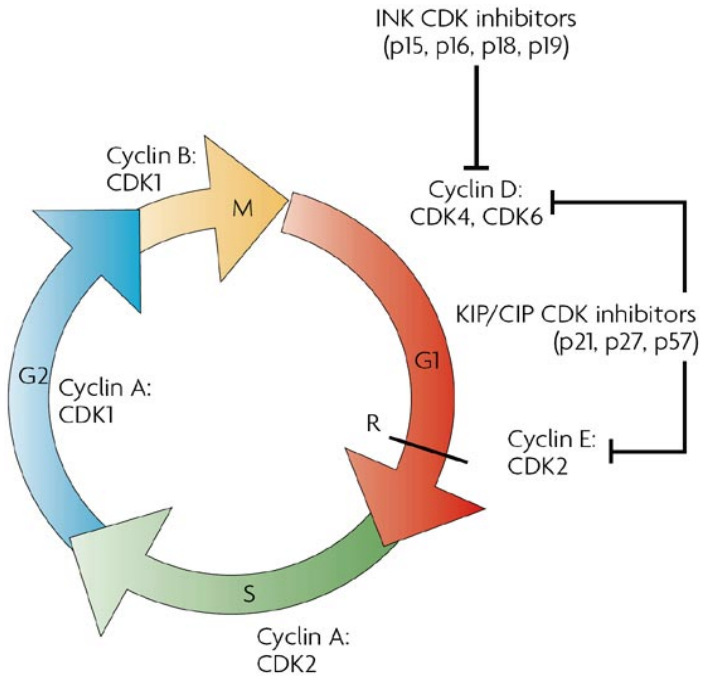


Figura 4. Control del ciclo celular.

Las distintas fases del ciclo celular junto a las ciclinas-CDKs, propias de cada estadío y los inhibidores de cada fase. Adaptado de: Colette Dehay et al. Nature Reviews Neuroscience (2007)

Este proceso contiene mecanismos estrictos de control que permiten la transición de una fase a otra una vez comprobada la integridad del material genético. De esta modo si algún error es detectado por las proteínas asociadas a las rutas DDR, éstas son capaces de desencadenar la parada del ciclo celular (Bartek and Lukas 2001, Bartek, Lukas et al. 2004). En consecuencia, la detención de la fase G1 se puede inducir con la expresión de la familia INK, entre ellas, INK4A (p16), INK4B (p15), INK4C (p18) e INK4D (p19), que inhiben CDK4 y CDK6. Además de forma alternativa, se pueden liberar proteínas de la familia KIP/CIP como p21, p27 y p57, que inhiben la actividad de CDK2 (Hanahan and

Weinberg 2011)(Figura 4). En último caso, se propagarían señales para la entrada en la muerte celular.

1.4.3. Replicación del ADN.

La replicación del ADN es un proceso primordial de la célula, que permite la duplicación del material genético, para su posterior transferencia a las células hijas durante la mitosis. Esta síntesis de ADN es llevada a cabo por las polimerasas. Actualmente, se conocen 15 tipos diferentes de las cuales la polimerasa α (Pol α), Pol δ y Pol ε son las encargadas de la replicación del ADN genómico del núcleo celular. El resto de ADN polimerasas desempeñan un papel crucial en la protección de la célula frente a daños en el ADN.

Debido a la elevada tasa de división de las células tumorales, se han identificado algunas polimerasas diferencialmente expresadas en varios tumores (Albertella, Lau et al. 2005). Por ejemplo, la sobre-expresión de *POLH* correlaciona con una peor prognosis en cáncer de pulmón (Ceppi, Novello et al. 2009), mientras que el aumento de la expresión de *POLB* es característico de carcinomas gástricos, uterinos, próstata, ovario y tiroides (Yoshizawa, Jelezcova et al. 2009). Por último, también se ha determinado la sobre-expresión de *POLQ* en tumores de colon (Kawamura, Bahar et al. 2004, Pillaire, Selves et al. 2010).

1.4.4. Elongación de los telómeros.

Los telómeros, localizados en los extremos de los cromosomas, están compuestos por repeticiones de hexa-nucleótidos, siendo su función principal garantizar la estabilidad estructural de los cromosomas en las células eucariotas. Estas secuencias de ADN no codificante altamente repetitivas, van disminuyendo de tamaño de forma progresiva con cada división celular, ya que la ADN polimerasa es incapaz de replicar las últimas 50-100 pares de bases del extremo 3'. De esta forma, el tamaño de las secuencias teloméricas determina el número total de divisiones celulares y por lo tanto el tiempo de vida de los tipos celulares.

Como se mencionó anteriormente, las telomerasas son las proteínas encargadas de la síntesis de las secuencias teloméricas y se caracterizan por no expresarse en las células normales. Sin embargo se expresan en la inmensa mayoría (~90%) de las células inmortalizadas, incluyendo las células tumorales (Blasco 2005). Su presencia está relacionada con el incremento de la proliferación celular, así como con la resistencia a la entrada tanto en senescencia como en la muerte celular. Sin embargo, la inhibición de la actividad telomérica que conlleva al acortamiento de los telómeros, provoca daños

genéticos y fusiones cromosómicas. Para evitar esto, existen mecanismos que se activan con la finalidad de desencadenar los procesos de senescencia o muerte celular.

1.4.5. Senescencia.

Senescencia es el termino que se empleó para definir el estadío celular no proliferativo que mostraban las células en medio de cultivo, descrito por Hayflick et al. hace más de cinco décadas (Hayflick 1965). Este proceso se caracteriza por la parada del ciclo celular en la fase G1, permitiendo un desarrollo celular completo pero a su vez, imposibilitando de forma irreversible, la capacidad de la célula de replicar su ADN y entrar en mitosis. Al igual que la apoptosis, la senescencia es una respuesta extrema al estrés celular y a mecanismos supresores de tumores (Green and Evan 2002). Sin embargo, la senescencia previene la división celular en aquellos casos donde aparecen alteraciones en el ADN genómico. La apoptosis, por el contrario, elimina la célula por completo. Asimismo, la mayoría de las células son capaces de activar ambas respuestas, pero se postula que la naturaleza y magnitud del daño en el ADN, pueden ser cruciales a la hora de activar una u otra ruta (Rebbaa, Zheng et al. 2003).

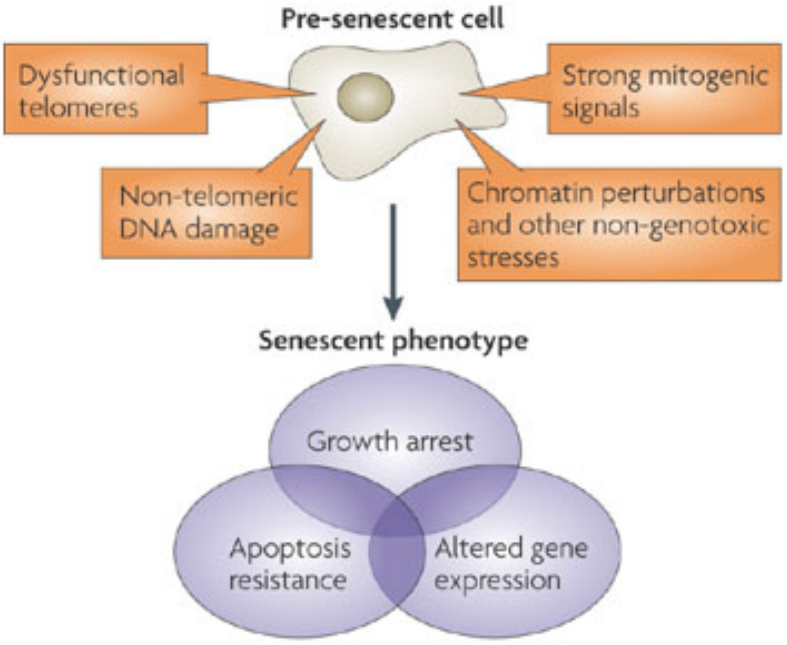


Figura 5. Fenotipo senescente dependiente de estímulo.

Estímulos relacionados con la entrada en senescencia, así como las características más representativas de las células senescentes. Adaptado de Campisi et al. Nature Reviews Molecular Cell Biology (2007).

Los estímulos descritos que desencadenan la senescencia son principalmente: la presencia de telómeros disfuncionales (normalmente telómeros suficientemente cortos como para permitir la división celular), daño severo en el ADN, generalmente asociado a roturas en la doble cadena del ADN (DSBs), alteraciones en la estructura organizativa de la cromatina (Bandyopadhyay, Okan et al. 2002) o sobre-expresión continuada de señales oncogénicas

de división celular (Zhu, Woods et al. 1998, Michaloglou, Vredeveld et al. 2005). Como resultado, las células senescentes se caracterizan, tal y como se muestra en la Figura 5, por la parada del ciclo celular y por consiguiente, por la sobre-expresión de inhibidores de proliferación celular como CDKIs p21 y p16 (Campisi 2001, Braig and Schmitt 2006).

1.4.6. Muerte celular.

La muerte celular entendida como un proceso de control que impide entre otras sucesos, la progresión tumoral, se puede diferenciar en dos mecanismos distintos:

1.4.6.1. Apoptosis.

La apoptosis fue descrita por primera vez por Kerr et al. y se define como el proceso de muerte celular principal en la célula que se induce especialmente ante la presencia de daños irreparables en el ADN (Kerr, Wyllie et al. 1972). En la apoptosis se suceden cambios morfológicos como la reducción del tamaño celular, la condensación y fragmentación del núcleo, así como la pérdida de la adhesión de la célula con células adyacentes y a la matriz extracelular (Nishida, Yamaguchi et al. 2008). A su vez, encontramos cambios bioquímicos que incluyen la rotura del ADN cromosómico en fragmentos (Cohen, Sun et al. 1994, Martin and Green 1995).

Existen dos rutas de activación del proceso apoptótico, la extrínseca y la intrínseca. La ruta extrínseca ocurre por la expresión de proteínas tipo FAS o BCL2, que forman parte del complejo de muerte inducido (DISC), activando una cascada de señalización en la que participan proteínas caspasas como CASP8, CASP9 y CASP3 (Kerr, Wyllie et al. 1972). Mientras que la ruta intrínseca, se caracteriza por la permeabilización y liberación del citocromo C de la mitocondria, lo que finalmente activa la red de las caspasas (Ghobrial, Witzig et al. 2005). Una vez que las caspasas han destruido el interior celular en forma de cuerpos apoptóticos, estos son digeridos por fagocitos sin ningún tipo de respuesta inflamatoria. Figura 6a.

1.4.6.2. Necrosis.

La necrosis se estimula mayoritariamente a través de los mediadores de muerte necrótica como PARP1, NADPH oxidasas y las calpaínas (Golstein and Kroemer 2007, Galluzzi and Kroemer 2008). Además, la activación de receptores de muerte o estrés celular puede a su vez inducir la activación de las quinasas RIP1 y RIP3. Esto implica a las mitocondrias, ya sea directamente o indirectamente a través de la NADPH oxidasa, a inducir el incremento de ROS, lo que conduce inevitablemente a la

necrosis (Kim, Morgan et al. 2007). También se ha descrito que las células durante el proceso de muerte por necrosis pierden la integridad de la membrana plasmática de tal forma que, los materiales intracelulares se liberan al medio extracelular, provocando una inflamación local y a su vez una posterior respuesta inflamatoria por parte de las células del sistema inmunitario. Figura 6b.

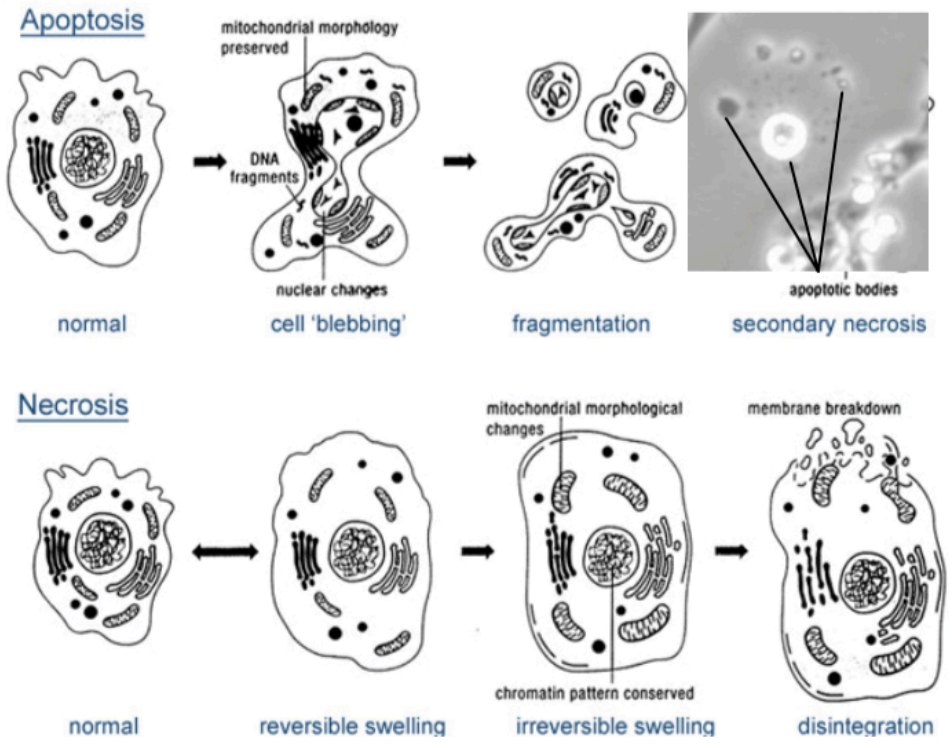


Figura 6. Tipos de muerte celular. a| Apoptosis. Señales procedentes de FAS o de la familia BCL desencadenan el desmantelamiento sistemático de la célula en forma de pequeños cuerpos apoptóticos, que son digeridos por los fagocitos sin inducir ninguna respuesta inflamatoria. **b| Necrosis**. Existe un aumento progresivo del tamaño celular hasta que la membrana plasmática se rompe y el contenido intracelular se libera al medio, provocando una respuesta inflamatoria. Adaptado de Mooma Hejmadi et al. 2010.

Un estudio mas profundo del equilibrio existente entre los niveles de expresión de los genes y su alteración en muestras tumorales, ya sean oncogenes o genes supresores de tumores implicados en estas rutas de regulación, es indispensable para el desarrollo de nuevas terapias efectivas contra el cáncer.

1.5. Papel de los microARNs en el cáncer.

Dentro de los diferentes reguladores génicos implicados en esta expresión, se encuentran los microARNs. Estos son moléculas de ARN de pequeño tamaño descubiertos hace más de dos décadas, capaces de regular el crecimiento celular, la diferenciación y la apoptosis,

entre una larga cantidad de procesos celulares. También, se han vinculado a importantes enfermedades como las afecciones cardiacas (Thum, Galuppo et al. 2007) o enfermedades neurodegenerativas (Nielsen, Lau et al. 2009). Además es conocido, que también pueden funcionar como oncogenes, llamados oncomires, o por el contrario, como supresores de tumores o anti-oncomires, y de esta forma, las modificaciones en los niveles de expresión de estos miARNs están estrechamente relacionados con la tumorogénesis.

1.5.1. miARNs: definición y origen.

Los microARNs se definen como ARNs mono-catenarios de una longitud media de 23 nucleótidos, que se transcriben a partir de genes de ADN no codificantes, es decir, no se traducen a proteínas. Los miARNs se identificaron por primera vez en 1993 por Lee y colaboradores en el laboratorio de Víctor Ambros (Lee, Feinbaum et al. 1993), durante un estudio realizado en *C. elegans,* sin embargo, poco después se demostró que los miARNs se expresaban en una amplia variedad de organismos, desde virus hasta plantas y humanos, destacando por su alta conservación entre especies (Bartel 2004).

La mayoría de los miARNs, ya sean intergénicos, intrónicos o policistrónicos, se transcriben mediante la actividad de la ARN polimerasa II, inicialmente como un transcrito primario o pri-miARN que presenta una caperuza en el extremo 5' y una cola de adeninas en el 3' (Poly-A). A continuación, estos pri-miARNs son procesados secuencialmente en en el núcleo celular por el complejo Drosha-DGCR8 generando una estructura de horquilla de aproximadamente 70 nucleótidos, denominada pre-miARN. Posteriormente, en el citoplasma, los pre-miARNs son procesados a miARNs maduros mediante la interacción con el complejo Dicer-TRBP (Figura 7). (Bernstein, Caudy et al. 2001, Grishok, Pasquinelli et al. 2001, Hutvagner, McLachlan et al. 2001, Lee, Ahn et al. 2003, Yi, Qin et al. 2003, Bohnsack, Czaplinski et al. 2004, Denli, Tops et al. 2004). Por último, estos miARNs maduros se incorporan en un complejo de silenciamiento inducido denominado miRISC en conjunto con la proteína Argonauta (AGO2), una ARNasa catalíticamente activa que se encarga de la degradación de los ARN mensajeros (ARNm) (Carthew and Sontheimer 2009, Krol, Loedige et al. 2010).

Actualmente, el número total de genes que codifican para miARNs es de 29000 para unos 200 organismos aproximadamente, según la última versión de la base de datos miRBase (Junio 2014). Concretamente, en el caso de humano, se han descrito 1900 genes, lo que implicaría que los miARNs podrían representar como mínimo el 10% del total de genes humanos (Kozomara and Griffiths-Jones 2014).

1.5.2. miARNs: función.

La función de los miARNs está relacionada con la regulación post-transcripcional de la expresión génica a través de la ruta de la ribo-interferencia (ARNi), (Pillai, Artus et al. 2004), mediante un emparejamiento de secuencia entre el microARN y el 3' UTR del ARN mensajero del gen.

Diversos estudios independientes han determinado que los miARNs regulan entre el 20-30% de los genes humanos, aunque algunos autores estiman que el porcentaje es muy superior, de aproximadamente un 92% (Lewis, Burge et al. 2005, Lim, Lau et al. 2005).

1.5.3. miARNs: mecanismos de acción.

A pesar del notable avance realizado en el conocimiento sobre el origen y la función de los miARNs, los mecanismos utilizados por estos para regular la expresión génica permanecen aún bajo un intenso debate. Concretamente, existen varios trabajos publicados que muestran que los miARNs en células animales pueden reprimir la expresión génica de hasta nueve maneras diferentes (Morozova, Zinovyev et al. 2012):

1. Inhibición de la unión del complejo ribosómico 40S para la formación del complejo de iniciación.
2. Inhibición de la unión del complejo ribosómico 60S.
3. Inhibición de la elongación de la traducción.
4. Finalización prematura de la traducción (disgregación de los ribosomas).
5. Degradación de la proteína durante la traducción.
6. Inhibición de la traducción mediante el secuestro del ARNm en los cuerpos-P (dominios citoplasmáticos que contienen proteínas implicadas en diversos procesos post-traduccionales, tales como la degradación del ARNm, la represión de la traducción y el silenciamiento de genes mediada por ARN).
7. Desestabilización del ARNm.
8. Degradación del ARNm.
9. Inhibición de la transcripción a través de la reorganización de la cromatina mediada por miARNs, seguida de un silenciamiento génico.

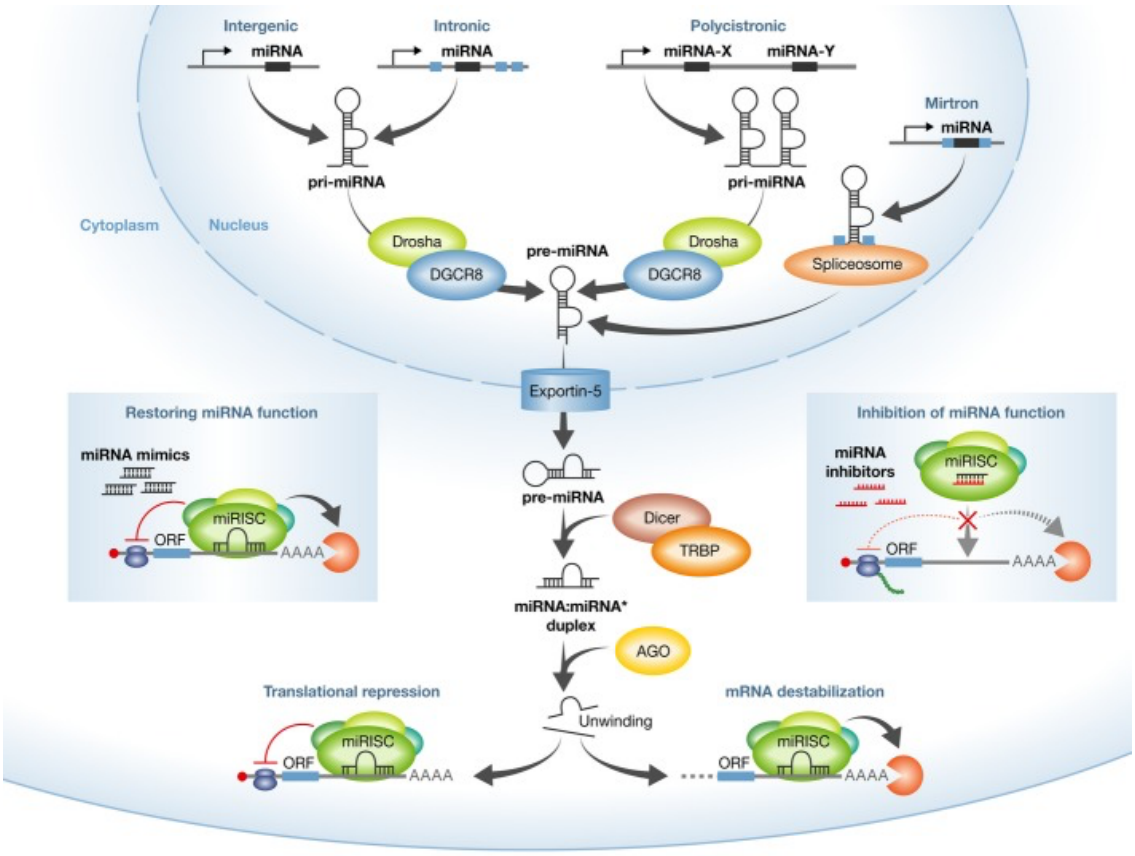


Figura 7. Origen y función de los miARNs. Los miARNs se transcriben por la polimerasa II generando un transcrito primario de gran tamaño. A continuación, este transcrito primario es procesado en el núcleo celular por el complejo Drosha-DGCR8 en forma de pre-miARN de 70 nucleótidos que es exportado al citoplasma a través de la exportina-5 donde junto con la proteína AGO2 forman el complejo de silenciamiento miRISC. Es en el citoplasma donde ejercen su función reprimiendo la traducción o desestabilizando el ARNm diana. Adaptado de Eva van Rooj. EMBO Molecular Medicine (2014).

Específicamente el emparejamiento de secuencia, viene determinado por los nucleótidos 2 al 7 de la parte 5' de los microARNs maduros (denominada región "semilla" del miARN), donde dichos nucleótidos deben ser complementarios con la región 3' UTR de uno o más ARNm. En plantas suelen tener un emparejamiento de secuencia miARN-ARNm completo o casi completo, induciendo el corte y la posterior degradación del ARNm diana (como ocurre con los siARNs en animales) (Saumet and Lecellier 2006). Sin embargo, en animales esta complementariedad de secuencia suele ser parcial, y generalmente los miARNs inhiben la traducción del ARNm o la degradación de los ARNm diana, mediante la eliminación de la caperuza en el extremo 5' y de la cola de poli-A en el extremo 3' (Eulalio, Rehwinkel et al. 2007). (Figura 7).

1.5.4. miARNs: relación con el cáncer.

En primer lugar, los miARNs fueron identificados por su capacidad de regular procesos asociados al desarrollo, tales como la diferenciación celular (Kloosterman and Plasterk 2006). Posteriormente, se les relacionó con la formación y progresión tumoral, así como con la metástasis (Shi, Liu et al. 2010, Wei, Wu et al. 2011). Varias líneas de investigación han elucidado que los miARNs se expresan diferencialmente en las células tumorales, donde crean un patrón de expresión único (Lu, Getz et al. 2005). Su desregulación está hoy en día reconocida como una propiedad común en los distintos tipos de cáncer humano (Figura 8).

Dentro de los miARNs que de forma extendida aparecen reprimidos en las muestras tumorales, encontramos a los supresores de tumores o anti-oncomires. Entre ellos resulta importante destacar a la familia let-7, considerada de las más antiguas y mejor conservadas, debido a su alta preservación de secuencia desde *Caenorhabditis elegans* hasta mamíferos (Pasquinelli, Reinhart et al. 2000). En humanos, la familia let-7 está compuesta por diez miembros (let-7a, let-7b, let-7c, let-7d, let-7e, let-7f, let-7g, let-7i, miR-98 y miR-202). Interesantemente, gran parte de los miARNs let-7 están ubicados en sitios frágiles relacionados con el cáncer (Calin, Dumitru et al. 2002). De tal forma, la perdida de la región 11q23, donde se alberga el clúster miR-125b1 / let-7a-2 / miR-100 es típica de los carcinomas de mama. Entre los genes regulados por esta familia, sobresalen oncogenes como *MYC*, un factor de transcripción nuclear relacionado con el ciclo celular, la apoptosis y la diferenciación y cuya sobre-expresión está relacionada con la formación de tumores (Denis, Kitzis et al. 1991). Igualmente, *HMGA2* (high-mobility grout AT-hook 2) es otro oncogen que regula la trascripción mediante cambios en la conformación de la cromatina (Peng, Laser et al. 2008).

Por el contrario, el clúster miR-17/92, un oncomiR presente en el cromosoma 13q31, se transcribe en un policistron que codifica para seis miARNs: miR-17, miR-18a, miR-19a, miR-20a, miR19b-1 y miR-92a-1. En comparación con los niveles de expresión de tejidos normales, este clúster aparece sobre-expresado en la mayoría de los tumores, como pulmón (Hayashita, Osada et al. 2005), linfomas de células B (Willimott and Wagner 2012), retinoblastoma (Sage and Ventura 2011), colon (Dews, Homayouni et al. 2006), páncreas (Morimura, Komatsu et al. 2011) y mama (Farazi, Horlings et al. 2011), entre otros. Así, por ejemplo, la región 13q31.3 donde este clúster se localiza en el genoma, está frecuentemente amplificada en linfomas difusos de células B grandes (Ota, Tagawa et al.

2004). Entre los genes regulados por este clúster, podemos resaltar la familia E2F como *E2F1* (Sylvestre, De Guire et al. 2007) y *E2F3* (Woods, Thomson et al. 2007), ambos desempeñan un papel esencial en la progresión del ciclo celular, específicamente en la transición de la fase G1 a la fase S. Así como el supresor de tumores *PTEN* (Ventura, Young et al. 2008) o la proteína P21 inhibidora de las CDK del ciclo celular (Fontana, Fiori et al. 2008).

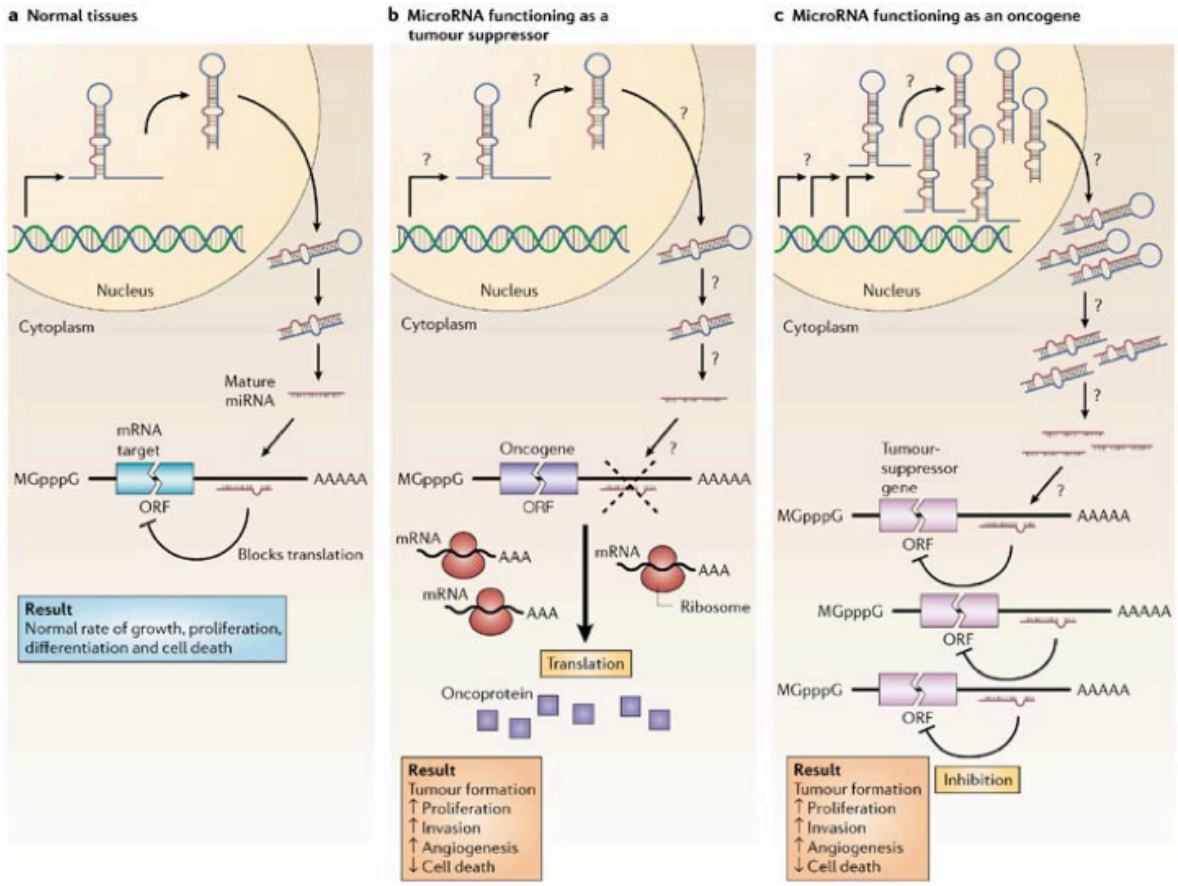


Figura 8. miARNs como supresores de tumores u oncogenes. a | En los tejidos normales, los miARNs se expresan y procesan para así reprimir la expresión de sus ARNm diana. El resultado global son tasas normales de crecimiento celular, proliferación, diferenciación y muerte celular. **b |** La reducción o eliminación de la expresión de un miARN supresor de tumores, puede ocurrir debido a defectos en cualquier etapa de la biogénesis del miARN (indicado por signos de interrogación), y en última instancia conduciría a la expresión inapropiada de distintitos genes, en numerosas ocasiones, oncogenes (cuadrados de color púrpura). El resultado global podría suponer un incremento de la proliferación, invasión o angiogénesis, disminución de la apoptosis, y en último lugar desembocaría en la formación de tumores. **c |** La amplificación o la sobreexpresión de un miARN oncogénico eliminaría la expresión de genes supresores de tumores (rosa) y provocaría la progresión del cáncer. Los elevados niveles de miARN maduro pueden ocurrir debido a la amplificación del gen miARN, un promotor constitutivamente activo o por un aumento de la estabilidad de los genes miARN (indicado por signos de interrogación). Adaptado de Esquela-Kerscher A et al. Nature Review Cancer (2006).

Tal es la implicación de los miARNs en el cáncer, que su estudio ha permitido el desarrollo de terapias contra el cáncer libre de sustancias químico-tóxicas. En 2005 Krützfeldt et al. generaron unos oligonucleótidos capaces de reprimir la función de diversos miARNs, denominándolos antagomirs (AMO). De esta forma, sintetizaron uno conjunto de oligonucleótidos con secuencia complementaria al miARN maduro, capaz de impedir la correcta formación del complejo miRISC y de esta forma evitando la degradación de los ARNm en una gran cantidad de tejidos (Krutzfeldt, Rajewsky et al. 2005).

Uno de los primeros casos de éxito de uso terapéutico de AMOS en cáncer, fue la desregulación del oncomiR miR-21. De tal forma que la inhibición de este miARN dio lugar a una disminución en la proliferación celular, acompañada de un aumento de procesos apoptóticos, tanto en líneas celulares de tumores de mama como en glioblastomas (Corsten, Miranda et al. 2007, Mei, Ren et al. 2010). En la actualidad encontramos diversos fármacos en fase I y fase II diseñados en base a AMOs o de forma contraria, en base a la mimetización de microARNs. Entre ellos destacamos, el antimiR-155 frente al linfoma de células T (Rupaimoole and Slack 2017) o el miR-16 frente a mesotelioma y tumor de pulmón no microcítico (Reid, Pel et al. 2013).

Las interacciones existentes entre los miARNs y los genes a los que regulan, son difíciles de estudiar experimentalmente. Los métodos actuales son complejos, caros y deben afrontar numerosos desafíos técnicos como son: la especificidad del tejido, la baja expresión, la selección de la 3' UTR a estudiar y, la estabilización del miARN (Huttenhofer and Vogel 2006), entre otros. En la actualidad existen diversas técnicas experimentales capaces de cuantificar la expresión del ARNm tras la co-expresión con un posible microARN regulador, como son los ensayos con luciferasa. Además, es posible cuantificar la cantidad de proteína derivada del ARNm a diferentes cantidades de un determinado miARN (Selbach, Schwanhausser et al. 2008), como son los ensayos basados en el marcaje de isótopos estables en aminoácidos dentro de cultivo celular (pSILAC).

1.5.5. Predicciones computacionales de interacciones entre miARN y ARNm.

Dado el alto nivel de complejidad de estas técnicas de laboratorio, las predicciones computacionales de interacciones miARN/ARNm emergen de forma complementaria para facilitar la caracterización experimental de estas asociaciones. Estos algoritmos se basan principalmente en la búsqueda de complementariedad de nucleótidos entre la secuencia semilla del miARN y alguna región del 3' UTR del gen, así como su termodinámica (Min and Yoon 2010). A continuación, tienen en cuenta otros factores obtenidos de interacciones

funcionales, como son: la conservación de la secuencia del sitio de unión entre distintos organismos (Farh, Grimson et al. 2005), el número de adeninas y uracilos junto al sitio de unión (Grimson, Farh et al. 2007), la presencia de abundantes sitios de unión cercanos para otros miARNs (Farh, Grimson et al. 2005) y una posición en el 3' UTR cercana a los extremos (Gaidatzis, van Nimwegen et al. 2007). En la actualidad, no existen algoritmos capaces de predecir interacciones en función del conjunto de todos los factores mencionados, de ese modo, se han creado diferentes bases de datos que almacenan predicciones proporcionadas por algoritmos diferentes. Entre estos, destacan miRonTop (Le Brigand, Robbe-Sermesant et al. 2010), miRGator (Cho, Jang et al. 2013), miRWalk (Dweep, Gretz et al. 2014) y MAGIA2 (Bisognin, Sales et al. 2012). Sin embargo, existe un escaso consenso entre las predicciones de los diferentes algoritmos de estos repositorios, además de una exigua fiabilidad de las predicciones cuando estas se intentan corroborar experimentalmente (Baek, Villen et al. 2008), de tal forma que resulta necesario la creación de una herramienta que sea capaz de aumentar la fiabilidad de las predicciones generadas.

Resolver este problema es primordial para mejorar la identificación de interacciones específicas y fiables entre microARNs y genes pertenecientes a las vías implicadas en el desarrollo del cáncer y además, susceptibles de desarrollo terapéutico. Para investigar si estas interacciones están implicadas en la tumorogénesis, es además, necesario el estudio tanto de genes cómo de microARNs directamente involucrados en la formación y proliferación del cáncer, como son aquellos que se obtienen tras un análisis de expresión diferencial, entre muestras de pacientes sanos y muestras tumorales. Para ello, es necesario el estudio común de muestras de transcriptómica tanto de ARN mensajeros (RNASeq) como microARNs (miRNASeq) de diferentes tipos de tumores, dentro de un marco estadístico conjunto.

Con el objetivo de aumentar la precisión de las interacciones entre genes y microARNs, en esta tesis doctoral se ha desarrollado una base de datos llamada miRGate (Andres-Leon, Gonzalez Pena et al. 2015), que permite obtener interacciones fiables a partir de genes y microARNs. Para estudiar la implicación de estos en la tumorogénesis, se ha desarrollado una segunda herramienta llamada miARma-Seq (Andres-Leon, Nunez-Torres et al. 2016) capaz de analizar un elevado número de muestras de expresión de genes y microARNs de forma integrada en diferentes tipos tumorales. Estos análisis podrían proporcionar información relevante sobre la estabilidad de estas interacciones en la biología del cáncer y su implicación en la supervivencia, para permitir el desarrollo de nuevas estrategias terapéuticas (Andres-Leon, Cases et al. 2017).

2. OBJETIVOS

El **objetivo principal** de esta tesis es desarrollar nuevos métodos computacionales, que permitan la identificación de redes de regulación con posible relevancia funcional entre genes y miARNs, que aparezcan de forma constitutiva desreguladas en pacientes diagnosticados de cáncer.

Para ello, se han establecido los siguientes **objetivos específicos**:

1. Definir una metodología capaz de predecir interacciones entre miARNs y ARNm con un alto valor de fiabilidad.

2. Desarrollar diferentes herramientas computacionales que permitan el estudio exhaustivo de una elevada cantidad de muestras de expresión de genes y miARNs procedentes de muestras tumorales de pacientes, así como muestras control.

3. Obtener un conjunto de genes y microARNs diferencialmente expresados entre muestras sanas y tumorales de pacientes.

4. Realizar un estudio global, para identificar aquellos genes y microARNs que aparecen constitutivamente desregulados en diferentes conjuntos de tumores, así como tumores de un mismo origen y las posibles interacciones de regulación entre ellos.

5. Analizar los resultados obtenidos eliminando posibles factores de confusión como la metilación y la alteración del número de copias, en la expresión génica.

6. Examinar el papel de las asociaciones obtenidas en los diferentes conjuntos de tumores, con la supervivencia de los pacientes.

3. MATERIALES Y MÉTODOS

En la presente sección se describe en detalle, la procedencia de la información y la metodología empleada en esta tesis doctoral. Además se especifican tanto los programas desarrollados por otros autores usados en esta tesis, así como los desarrollados personalmente durante este trabajo para la obtención de los resultados.

3.1. Interacciones miARN-ARNm.

Para acometer el objetivo principal de esta tesis, es necesario analizar un elevado número de muestras de pacientes y así, obtener genes y microARNs diferencialmente expresados que permitan estudiar los fenómenos de regulación existentes entre ellos. En la actualidad, para investigar la regulación llevada a cabo por los microARNs, se han creado múltiples algoritmos que identifican sitios de unión potenciales a disposición de los investigadores. Entre estos algoritmos, es conocido el poco consenso existente entre sus predicciones y la escasa fiabilidad de estas interacciones potenciales, al verificarlas experimentalmente (Baek, Villen et al. 2008). Con el objetivo de eliminar o reducir estos obstáculos, se ha desarrollado miRGate (Andres-Leon, Gonzalez Pena et al. 2015).

3.1.1. miRGate.

miRGate es una base de datos que contiene predicciones miARN-ARNm para humano, rata y ratón. Se basa en la integración de las predicciones obtenidas a partir de métodos distintos, dado que es la forma más eficaz de obtener interacciones precisas en función de los diversos factores implicados en las asociaciones funcionales conocidas (Ritchie, Flamant et al. 2009). Sin embargo como se ha comentado, los diferentes algoritmos disponibles, presentan una gran limitación, como es la variabilidad de sus predicciones, incluso entre algoritmos cuyos cálculos se fundamentan en factores similares. Tal y como resume la Tabla 1, programas como miRanda (Betel, Koppal et al. 2010), Pita (Kertesz, Iovino et al. 2007) y Pictar (Krek, Grun et al. 2005) usan versiones genómicas distintas para las secuencias 3’ UTRs y para los microARNs, lo que genera una escasa similitud en los resultados obtenidos.

Al contrario, nuestra base de datos miRGate, en lugar de partir de interacciones pre calculadas, usa un conjunto común de secuencias procedentes de las 3’UTR de los genes y de las secuencias de miARNs para calcular nuevas interacciones a través de cinco métodos diferentes. A continuación se detalla el conjunto de secuencias y los algoritmos de predicción utilizados en miRGate.

3.1.2. Secuencias de miARNs y ARNms.

Todas las secuencias 3' UTR del genoma humano versión GRCh37.p13, de ratón versión GRCm38.p2 y de rata versión Rnor_5.0 se descargaron desde EnsEMBL 74 (Yates, Akanni et al. 2016). A diferencia de otras bases de datos, miRGate no solo usa las secuencias de isoformas que codifican proteínas, si no que también incorpora todas aquellas isoformas que puedan expresarse, como pseudogenes, relacionados con la actividad reguladora de genes involucrados en cáncer (Poliseno, Salmena et al. 2010), isoformas que retienen intrones o genes que codifican para inmunoglobulinas, entre otras y en resumen los 17 tipos distintos de isoformas recogidos por el proyecto HAVANA en su página web: http://vega.sanger.ac.uk/info/about/gene_and_transcript_types.html. En la Tabla 1a se muestra la comparativa entre las secuencias 3'-UTR en humano incluidas en miRGate, frente a otras bases de datos.

Tabla 1. Conjunto de secuencias 3'UTR y de microARNs usados en miRGate y otros.

a| Conjunto de secuencias 3' UTR incluido en miRGate frente a otros.

Nombre	Genoma | Año	Genes Codificantes	Genes No Codificantes	Pseudogenes	3' UTR
miRanda	NCBI37 2009	19.778	-	-	34.592
Targetscan	NCBI37 2009	18.414	-	-	30.932
Pita	NCBI36 2006	18.582	-	-	24.086
PicTar	NCBI35 2005	20.254	-	-	20.254
miRGate	NCBI37 2009	20.805	22.966	14.181	196.501

b| Conjunto de secuencias de miARNs incluidos en miRGate frente a otros.

Nombre	Humano	Ratón	Rata	Versión
miRanda	1.1	717	387	miRBase 15
Targetscan	1.433	722	-	miRBase 17
Pita	692	500	-	miRBase 11
PicTar	81	81	81	Rfam 5
miRGate	2.68	1.983	763	miRBase 20

En el caso de los microARNs, las secuencias se obtuvieron de miRBase versión 20 (Kozomara and Griffiths-Jones 2014), la principal base de datos existente, para estos ARNs. Específicamente, se obtuvieron las secuencias de todos los microARNs que potencialmente podrían regular genes humanos, incluyendo no solamente miARNs de humano, sino también de virus patógenos, como Epstein-Barr y el citomegalovirus. Además de la información completa para humano, miRGate asimismo, contiene los

microARNs descritos tanto en rata cómo en ratón. El número total de miARNs de las distintas especies incluidas en miRGate y en el resto de bases de datos se indica en la Tabla 1b.

3.1.3. Algoritmos.

Actualmente existen numerosos programas que proporcionan interacciones entre miARNs y genes, que han sido calculadas a partir de una secuencia del gen (por lo tanto sólo una isoforma por cada gen). Sin embargo, los programas que dada una secuencia cualquiera de una región 3' UTR de un gen y la secuencia de un miARN, permitan calcular una interacción, son escasos. En consecuencia, miRGate sólo ha utilizado aquellos programas que posibilitaran el calculo de nuevas interacciones a partir de un conjunto de secuencias de nuestra elección, como son:

(i) **miRanda** (Enright, John et al. 2003) es un método de predicción en tres fases. En primer lugar, evalúa la complementariedad de la secuencia entre los elementos a interaccionar, después realiza el cálculo de la energía libre para estimar si la interacción física es favorable, y por último, analiza la conservación del sitio de unión entre organismos cercanos evolutivamente, en nuestro caso concreto, se utilizaron los datos de humano, rata y ratón.

(ii) **Pita** (Kertesz, Iovino et al. 2007), programa caracterizado por presentar como requisito imprescindible en el sitio de unión, la complementariedad completa de la secuencia formada por la región semilla de los miARNs y las secuencias 3' UTR proporcionadas. Además, lleva a cabo el cálculo de energía libre entre la doble cadena unida de ARNs y las cadenas simples separadas, para verificar uniones termodinámicamente propicias. Por último, con el objetivo de filtrar todas aquellas predicciones poco conservadas, este programa es capaz de incorporar datos de conservación. Particularmente, nosotros proporcionamos los datos de 46 genomas de vertebrados calculados mediante Phastcons (Siepel, Bejerano et al. 2005), un programa filogenético basado en modelos de Markov.

(iii) **RNAHybrid** (Kruger and Rehmsmeier 2006), método capaz de generar predicciones energéticamente favorables sobre múltiples sitios de unión potenciales entre secuencias de ARNs de tamaño diverso. En general, el programa identifica aquellos sitios de hibridación energéticamente más favorables, entre un ARN de pequeño tamaño frente a ARN de mayor tamaño. Asimismo, utiliza aproximaciones de Poisson para el cálculo de numerosos

sitios de unión y filtros de energía para evaluar los resultados obtenidos.

(iv) **TargetScan** (Friedman, Farh et al. 2009) se basa en predecir interacciones fundamentadas en la complementariedad completa de la secuencia, la localización del sitio de unión a lo largo de la secuencia del ARNm, el enriquecimiento de adeninas y uracilos, y la conservación evolutiva del sitio de unión entre especies distintas. Específicamente para este algoritmo, se emplearon los alineamientos de secuencia EPO (Enredo, Pecan/Ortheus) proporcionados por la base de datos EnsEMBL para el conjunto de mamíferos que integran.

(v) **Microtar** (Thadani and Tammi 2006) es un algoritmo que permite el cálculo de interacciones miARN-ARNm basadas en una complementariedad incompleta de la secuencia en el sitio de unión. Igualmente, estima la energía del dúplex de ARN para la obtención de predicciones energéticamente posibles.

Toda la información proporcionada por estos algoritmos, como son: el tipo de sitio de unión, los valores de energía y de conservación evolutiva, así como las coordenadas genómicas de la zona de unión en la UTR, fue almacenada en una base de datos relacional en MySQL. El flujo de trabajo completo del funcionamiento de miRGate se presenta en la Figura 9.

3.1.4. Conjunto de datos validados experimentalmente.

Con el propósito de obtener las predicciones más precisas posibles, y a su vez, poder destacar las interacciones obtenidas y almacenadas dentro de la base de datos miRGate, se recopilaron todas aquellas uniones miARN-ARNm verificadas experimentalmente y obtenidas a partir de cuatro repositorio públicos, como son:

(i) **Tarbase** (Vergoulis, Vlachos et al. 2012) es una base de datos que contiene amplia información sobre cada interacción miARN-gen publicada en Pubmed. Además, mediante el uso de técnicas de minería de datos, almacena información específica sobre la interacción, incluyendo las metodologías empleadas para su validación. Todas las entradas de esta base de datos engloban también información general derivada de distintas fuentes externas, tales como UniProt, EnsEMBL y RefSeq.

(ii) **miRTarbase** (Hsu, Tseng et al. 2014) de forma similar, aplica herramientas de minería de datos para identificar aquellas interacciones verificadas empíricamente entre los resúmenes de los artículos publicados en PubMed.

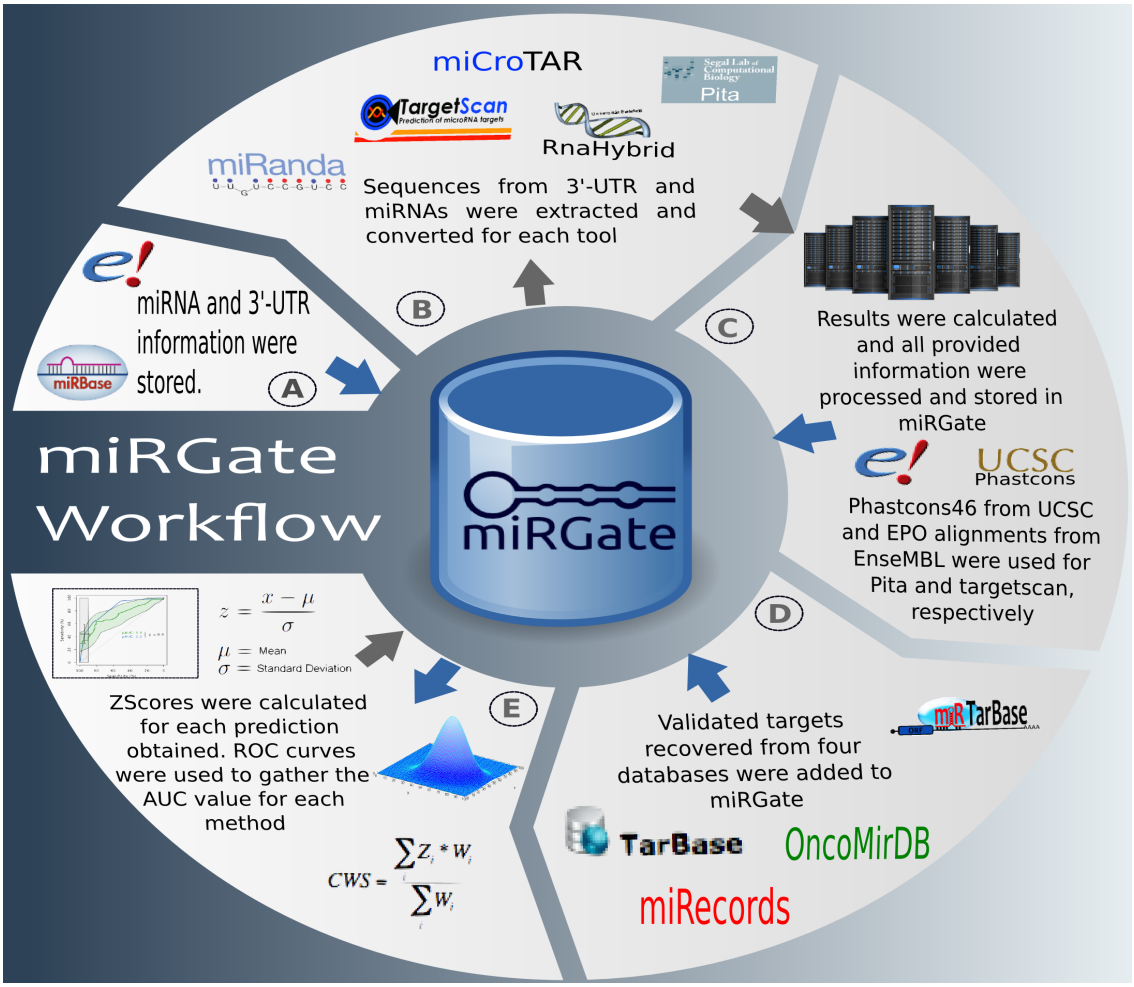


Figura 9. Representación del flujo de trabajo de miRGate. a| El conjunto de secuencias e información adicional de miARNs es obtenida automáticamente de la versión 20 de miRBase y en el caso de los genes, de la versión 74 de EnsEMBL para ser almacenada en una base de datos relacional en MySQL. b| Esta información es procesada según los formatos de entrada de cada algoritmo (miRanda, RNAHybrid, Pita, microTar y Targetscan). c| Estos programas son ejecutados en ordenadores de alto rendimiento. La información adicional necesaria, así como la información de conservación de Phastcons para Pita o de EPO para Targetscan, también es incluida en este proceso. Los resultados obtenidos, como son los valores de puntuación o scores, valores de energía, conservación y coordenadas de unión son procesados e insertados en la base de datos. d| La información sobre las interacciones validadas experimentalmente es añadida en este paso. e| Los valores de puntuación generados en cada predicción son estandarizados y, posteriormente dichos valores son examinados frente a datos validados para obtener curvas ROC y valores de fiabilidad.

(iii) **miRecords** (Xiao, Zuo et al. 2009) es una base de datos que contiene información obtenida manualmente tras la evaluación de un extenso volumen de artículos científicos, que han sido seleccionados automáticamente por presentar información sobre posibles dianas de regulación entre miARNs y genes.

(iv) **OncomirDB** (Wang, Gu et al. 2014) reúne interacciones validadas experimentalmente que han sido obtenidas a partir de aproximadamente 9.000

artículos científicos publicados en relación con el cáncer. Asimismo, toda la información es verificada manualmente y los datos experimentales son almacenados.

Por último, resulta importante destacar que miRGate proporcionó diversas predicciones que fueron validadas posteriormente en sucesivas colaboraciones por diferentes laboratorios. Entre ellas sobresalen los trabajos realizados junto al Dr. Javier Benítez en el Centro Nacional de Investigaciones Oncológicas (CNIO) en cáncer hereditario de mama (Tanic, Andres et al. 2013) o en cáncer de triple negativo de mama (Matamala, Vargas et al. 2016). Adicionalmente, con el Dr. Miguel Ángel Piris del Hospital Universitario Marqués de Valdecilla de Santander, participamos en los estudios sobre linfoma de células del manto (Di Lisio, Gomez-Lopez et al. 2010), linfomas de células B (Di Lisio, Martinez et al. 2012) y linfomas difusos de células B asociados al virus Epstein-Barr (Martin-Perez, Vargiu et al. 2012). En otra colaboración, en este caso llevada a cabo con el laboratorio del Dr. Javier Cigudosa perteneciente también al CNIO, se validaron exitosamente predicciones para mielomas múltiples híper-diploides (Rio-Machin, Ferreira et al. 2013). Finalmente, junto con el grupo de la Dra. Giuglia de Falco de la Universidad de Siena, se ratificaron experimentalmente diversas predicciones en linfomas de Burkitt asociados al virus Epstein-Barr (Ambrosio, Navari et al. 2014).

3.1.5. Z-score y concordancia genómica.

El conjunto de predicciones obtenido para cada organismo, es ordenado en base a un valor de Z-score. Este Z-score es calculado al estandarizar cada valor de puntación o *score* generado para cada predicción y proporcionado por cada algoritmo. Para ello, a cada valor se le substrae la media del conjunto de predicciones y a continuación se divide el resultado entre la desviación estándar. En el caso de que una misma predicción, y por lo tanto, en la misma coordenada genómica, fuera obtenida por más de un método distinto, obtendríamos lo que hemos denominado "concordancia genómica". Los valores de las predicciones con concordancia genómica se combinan usando un valor pesado y ponderado, CWS (del inglés *Consensus Weighted Score*) previamente utilizado por otros autores (Gonzalez-Perez and Lopez-Bigas 2011) y que se define en la siguiente formula:

$$CWS = \frac{\sum_i Z_i * W_i}{\sum_i W_i}$$

De esta manera, por cada predicción idéntica obtenida por un determinado algoritmo, definimos Z_i como el valor estandarizado generado por ese algoritmo, y siendo W_i, la probabilidad de que una predicción no sea un falso positivo, dado la distribución complementaria acumulativa de valores demostrada por la herramienta *i* cuando se comparan sus predicciones frente a una conjunto de interacciones validadas.

Una vez que la base de datos de predicciones fue creada, para avanzar al siguiente objetivo específico de la tesis doctoral, necesitábamos poder obtener un listado de genes y microARNs diferencialmente expresados procedentes del análisis de muestras tumorales. Esto se llevo a cabo mediante la creación de una herramienta para el procesamiento de muestras de expresión obtenidas mediante técnicas de Secuenciación de Nueva Generación (NGS), denominada miARma-Seq (Andres-Leon, Nunez-Torres et al. 2016).

3.2. Análisis computacional de las muestras de miARNs y ARNms.

Estas muestras de transcriptómica, obtenidas por técnicas de NGS se caracterizan por diferir considerablemente según el tipo de datos de expresión a estudiar. Los datos de expresión de genes, derivados de RNA-Seq, presentan lecturas originadas de la secuenciación de ambos extremos de un fragmento (*paired-end*) y con un tamaño de secuencia variable entre 75 y 150 nucleótidos. Sin embargo, los datos provenientes de miRNASeq se caracterizan por la secuenciación de sólo un extremo (*single-end*) y tener un tamaño de secuencia inferior, concretamente alrededor de 50 nucleótidos. Por tanto, los algoritmos empleados para su estudio y análisis, son diferentes. En esta tesis doctoral se ha desarrollado una herramienta llamada miARma-Seq (Andres-Leon, Nunez-Torres et al. 2016) con el objetivo de poder analizar un número muy elevado de muestras de forma simultánea, atendiendo a las características propias de los ARNm y miARNs y solventando algunas de las limitaciones que imponen las herramientas disponibles en la actualidad.

3.2.1. miARma-Seq.

miARma-Seq ó del ingles *miRNA and mRNA multiprocess analysis,* es una nueva herramienta que permite realizar un análisis en profundidad a partir de muestras de transcriptómica, ya sean miARNs, genes o ARNs circulares (circARNs). Este programa implementa un conjunto de libreras en Perl y R. Para facilitar su uso, miARma puede ejecutarse como un simple comando de consola junto a un fichero de configuración, que contiene información referente al experimento de NGS. Sin embargo, miARma-Seq también puede ser usado por personas con experiencia en programación ya que las librerías

se encuentran disponibles de forma abierta para su utilización e integración con otros flujos de trabajo a través del siguiente link https://bitbucket.org/cbbio/miarma/src.

A continuación se describen los aspectos más relevantes de miARma-Seq.

3.2.2 Características principales de miARma-Seq.

Esta herramienta se caracteriza por integrar internamente todos los programas necesarios para llevar a cabo un análisis de muestras de NGS, reduciendo de esta forma, el número de dependencias, al mínimo. A diferencia de otros programas, permite el análisis completo y simultáneo de muestras de NGS desde los datos brutos (Figura 10) de manera sencilla. Además, debido a su diseño modular, ofrece la flexibilidad de iniciar un análisis en cualquier punto del proceso, en el caso de tener resultados parciales obtenidos por otras herramientas alternativas. Esta particularidad de miARma-Seq, resulta crucial cuando se usan datos obtenidos de bases de datos publicas como son GEO o SRA. Adicionalmente, aunque se suministran opciones configuradas para un análisis por defecto que incluyen diversos parámetros ajustados según en análisis seleccionado (miARN, ARNm o circARN), el fichero de configuración posibilita al usuario elegir el software y los parámetros específicos más adecuados para analizar sus datos. Este software integrado internamente en miARma, ha sido seleccionado mediante la comparación y evaluación del conjunto de herramientas más usadas en el campo del análisis de RNASeq y miRNASeq (Nookaew, Papini et al. 2012, Williamson, Kim et al. 2013, Fonseca, Marioni et al. 2014, Fan, Zhang et al. 2015) o por ofrecer importantes ventajas para los análisis (predicción de secuencias adaptadoras, simulación de replicados, entre otras).

A continuación se exponen los pasos más habituales en un análisis de datos de expresión procedentes de NGS, destacando el software disponible en miARma-Seq para tal efecto.

- **Análisis de calidad y pre-procesamiento.** Inicialmente y siempre que se disponga de ficheros en formato fastq, las muestras pueden analizarse con el objetivo de detectar errores de secuenciación para por ejemplo, el posterior descarte de secuencias de baja calidad. Con este propósito, miARma incluye FastQC (Anders 2010). Este programa proporciona un informe riguroso sobre la calidad media de los nucleótidos a lo largo de las lecturas secuenciadas. Asimismo, permite comprobar la inclusión excesiva de adaptadores en las lecturas (debido a la baja complejidad de las secuencias en los extremos) o el porcentaje de secuencias duplicadas. De esta forma, y para eliminar estos nucleótidos de baja complejidad, adaptadores y/o nucleótidos de baja calidad, miARma-Seq incorpora diferentes

programas, entre ellos destacamos Cutadapt (Creighton, Nagaraja et al. 2008) o Reaper (Davis, van Dongen et al. 2013). Igualmente, es conveniente resaltar que los miARNs maduros presentan un tamaño estimado de 22 nucleótidos y dado que el tamaño de las lecturas es superior en la mayoría de los casos, el pre-procesamiento de las muestras es obligatorio debido a que la presencia de secuencias adaptadoras podría interferir en la correcta alineación de las lecturas y, en particular, en la de secuencias de un tamaño inferior a 50 nucleótidos. Cutadapt y Reaper están diseñados para eliminar la secuencia adaptadora o parte de ella, así como para llevar a cabo un filtrado adicional de baja complejidad. Un problema frecuente cuando se trabaja con datos obtenidos de repositorios públicos, es la falta de información experimental disponible, incluyendo la naturaleza de las secuencias adaptadoras, información esencial para ejecutar este proceso correctamente. Para solventar esta limitación, miARma-Seq implementa Minion (Davis, van Dongen et al. 2013), programa que realiza una predicción de la secuencia adaptadora en base a los nucleótidos más frecuentes en el extremo 3' de la lectura. En este sentido, miARma-Seq es capaz de predecir de forma automática la secuencia adaptadora incluso si no se proporciona esta información adicional. Además, nuestra herramienta realiza una comprobación de la secuencia obtenida, con el fin de descartar secuencias biológicas altamente representadas y que pudieran ser identificadas como adaptadores potenciales. Para ello, miARma-Seq utiliza la herramienta Blat de la Universidad de Santa Cruz en California (https://genome.ucsc.edu/cgi-bin/hgBlat) para diferenciar adaptadores frente a secuencias biológicas. De esta manera, con la incorporación de Minion, se pueden obtener las secuencias adaptadoras para el posterior análisis con Cutadapt o Reaper. Por último, destacar que miARma-Seq también presenta una utilidad propia, que permite eliminar un número específico de nucleótidos desde el extremo 3' o 5' por contener estos en lo general, nucleótidos de baja calidad.

- **Alineamiento.** Este proceso consiste en localizar en el genoma de referencia en base a su complementariedad de secuencia, cada una de las lecturas proporcionadas en el fichero fastq. Como resultado de esta fase, se genera un fichero que además de contener la secuencia de origen, incluye las coordenadas genómicas en relación al lugar en el que la secuencia se encuentra localizada. Este paso se diferencia según la naturaleza de los datos a analizar. En el caso de los miARNs, el alineador por

defecto es Bowtie1 (Langmead, Trapnell et al. 2009), ya que es el alineador con mejor sensibilidad y mejor rendimiento para secuencias de tamaño inferior a 50 nucleótidos, aunque si el usuario quiere, puede usar Bowtie2 (Langmead and Salzberg 2012) o incluso ambos (lo que conlleva el doble alineamiento del conjunto de lecturas con un software y luego con el otro). Por otra parte, miARma-Seq proporciona la capacidad de predecir miARNs no conocidos. De esta forma, se incluye mirDeep2 (Friedlander, Mackowiak et al. 2012), un programa que se basa en la utilización de aquellas lecturas no alineadas frente al genoma de referencia y que puedan ser usadas para predecir nuevos miARNs en base a diferentes características, como la capacidad de un pre-miARN para formar una estructura de horquilla o seguir el modelo de procesamiento de la proteína Dicer.

Para el alineamiento de ARNm, miARma-Seq emplea TopHat (Kim, Pertea et al. 2013), programa que a parte de alinear la secuencias de partida, es capaz de detectar la expresión de isoformas y genes fusionados. TopHat está formado por un conjunto de programas que permite alinear las lecturas mediante el uso de Bowtie2 (por defecto), o si se quiere con Bowtie1 o incluso ambos.

- **Cuantificación de lecturas.** Tras posicionar las secuencias en el genoma de referencia, es posible cuantificar el número total de secuencias alineadas por cada gen, dado que en un análisis de Expresión Diferencial (DE), este número es proporcional a la expresión del gen en el experimento. Con este objetivo, se ha incluido FeatureCounts (Liao, Smyth et al. 2014), un programa capaz de cuantificar y proporcionar el número de lecturas por cada gen o miARN.

- **Análisis de expresión diferencial.** Una gran parte de los estudios de expresión se centran en la identificación de entidades con niveles de expresión estadísticamente alteradas entre dos condiciones experimentales. Existen numerosos programas para la identificación de genes y/o miARNs que hayan sufrido cambios significativos de sus niveles de expresión, concretamente, miARma-seq incluye edgeR (Robinson, McCarthy et al. 2010) y NOISeq (Tarazona, Garcia-Alcalde et al. 2011). En el caso de edgeR, se trata de un software desarrollado en R y ampliamente usado en la comparación de muestras con dos condiciones experimentales. Pero, a su vez permite ser utilizado en comparaciones más complicadas. Por lo tanto, a diferencia de la mayoría de las herramientas analíticas disponibles, que sólo aceptan comparaciones básicas de dos condiciones experimentales, miARma-Seq mediante

la incorporación de edgeR, identifica elementos diferencialmente expresados derivados de cualquier tipo de diseño experimental, ya sean series temporales o efectos combinados debido al uso de medicamentos.

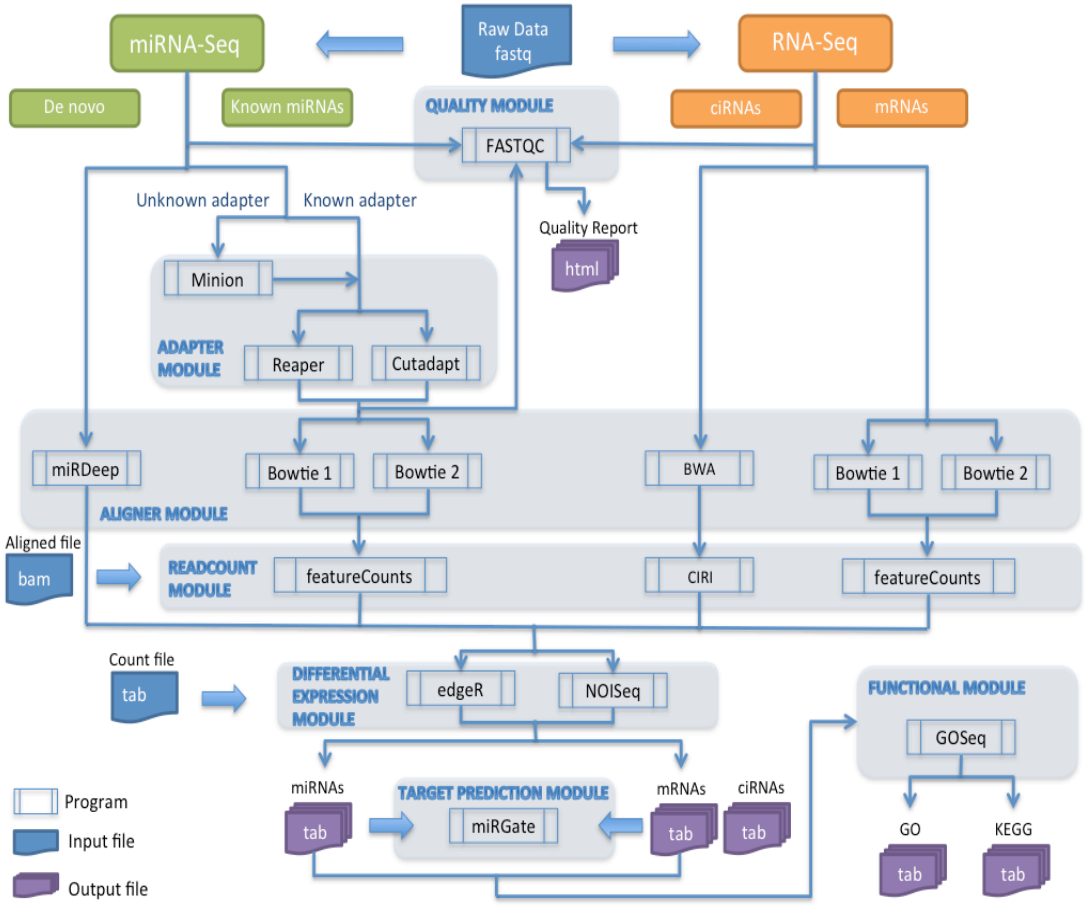


Figura 10. Gráfico resumen del diseño modular de miARma-Seq. Los módulos principales se muestran con un color de fondo gris.

Otra limitación habitual en los análisis de expresión diferencial, es la falta de réplicas biológicas o técnicas para cometer un análisis estadístico fiable. Aunque esta situación está totalmente desaconsejada a los usuarios debido a la escasa fiabilidad de los resultados, las dificultades experimentales como la falta de muestras o la baja calidad de las réplicas, en ocasiones obligan a los investigadores a efectuar sus ensayos en ausencia de replicados. Para solventar estas condiciones, miARma-Seq implementa NOISeq, este programa posibilita la simulación tanto de réplicas biológicas como técnicas, con la finalidad de incrementar la solidez de los resultados. También, otro factor destacable que afecta a la fiabilidad de los resultados, es la presencia tanto de genes como miARNs con muy baja expresión, por este motivo, miARma incluye la opción de eliminar aquellos elementos de

escasa expresión. Este alternativa puede ser elegida por el usuario, aunque por defecto tal y como recomienda Anders y colegas (Anders, McCarthy et al. 2013), miARma descarta, todos los genes/miARNs que se expresen por debajo de una lectura por cada millón de lecturas alineadas (CPM, del inglés *Counts Per Million)* en el conjunto de réplicas. Por último, múltiples parámetros, como el algoritmo a emplear para la normalización del grupo de muestras, pueden ser seleccionados por el usuario.

- **Predicción de las interacciones miARN-ARNm.** La mayoría de los programas que analizan muestras de transcriptómicas que han sido obtenidas a través de técnicas de NGS, suelen proporcionar un listado de genes o miARNs diferencialmente expresados como resultado final. Sin embargo, miARma-Seq ofrece la posibilidad de predecir posibles interacciones entre los miARN y los ARNm diferencialmente expresados. En el caso de un análisis basado en la expresión diferencial de genes, nuestra herramienta proporciona los posibles miARNs reguladores. Por el contrario, en el caso de estar analizando miARNs estadísticamente desregulados, miARma, es capaz de aportar posibles dianas génicas sujetas a su regulación. Sin embargo, si el experimento comprende datos de transcriptómica, tanto de miARNs como de genes, este módulo realiza el calculo automático de interacciones, entre perfiles de expresión invertidos, es decir miARNs reprimidos frente a ARNm sobre-expresados, y viceversa. Para llevar a cabo estas predicciones, miARma-Seq emplea miRGate (Andres-Leon, Gonzalez Pena et al. 2015) una herramienta desarrollada también en este trabajo y que se detalla en el apartado 3.1.1.

- **Análisis Funcional.** Adicionalmente, en el caso de un análisis de expresión diferencial de genes, miARma-Seq puede realizar un estudio de procesos y rutas metabólicas alteradas estadísticamente, mediante el análisis del conjunto de genes desregulados del experimento. Para llevar esto a cabo, miARma-Seq emplea GOSeq (Young, Wakefield et al. 2010), que permite hacer un estudio de los genes asociados al complejo celular, proceso biológico y función molecular, junto a las rutas metabólicas procedentes de la base de datos KEGG (Kanehisa, Sato et al. 2016). Estos análisis funcionales son de gran utilidad, ya que permiten abordar en profundidad el conocimiento de los procesos biológicos alterados, a través de los genes implicados.

3.3. Muestras transcriptómicas de pacientes.

Con el propósito de identificar redes de regulación conservadas en cáncer, resulta primordial, el acceso a un amplio número de muestras tumorales y muestras control. Concretamente, en este trabajo se recurrió al conjunto de datos procedentes del Atlas Genómico del Cáncer (TCGA, del inglés, *The Cancer Genome Atlas*). Las muestras empleadas en este trabajo, proceden de datos crudos (ficheros en formato fastq o BAM) de acceso restringido. Por lo tanto, para la obtención de estas muestras, se solicitó permiso al Instituto Nacional del Cáncer (NCI) y al Instituto de Investigación de Genómica Humana (NHGRI), ambos pertenecientes al Instituto Nacional de Salud Americano (NHI). En la actualidad, el TCGA engloba muestras provenientes de distintas plataformas de 33 tipos de tumores diferentes. Dado que para llevar a cabo nuestro análisis, se requiere del estudio simultáneo de muestras de transcriptómica de genes y de microARNs, todos aquellos tumores que no poseían este tipo de información, fueron descartados. Posteriormente, de los 19 tipos de tumores restantes y con el fin de poder realizar un estudio estadístico exhaustivo de pacientes sanos y enfermos, se excluyeron los tipos de tumores con menos de 10 muestras sanas o los tumores cuyo número total de muestras sanas no llegara al 5% del total. Tras aplicar este filtro, se obtuvo un total de 15 tipos de tumores distintos:

1. Tumor de riñón cromófobo (KICH).
2. Carcinoma escamoso de cuello y cabeza (HNSC) .
3. Carcinoma esofágico (ESCA).
4. Carcinoma de riñón de célula papilar (KIRP).
5. Carcinoma hepático (LIHC).
6. Carcinoma de riñón de célula clara (KIRC).
7. Adenocarcinoma de pulmón (LUAD).
8. Carcinoma de tiroides (THAD).
9. Adenocarcinoma de próstata (PRAD).
10. Carcinoma urotelial de vejiga (BLCA).
11. Carcinoma invasivo de mama (BRCA).
12. Carcinoma escamoso de pulmón (LUSC).
13. Adenocarcinoma de estómago (STAD).
14. Cholangiocarcinoma (CHOL).
15. Carcinoma endometrial de Útero (UCEC).

El TCGA contenía un total de 18.605 muestras (6.867 RNASeq y 11.738 miRNASeq) de los 15 tipos de tumores seleccionados. Estas muestras se fueron descargando y analizando de forma progresiva desde el día 21 de diciembre de 2015. Un resumen del número de muestras por tumor se detalla en la Tabla 2.

		RNASeq		miRNASeq	
Tipo de tumor	**Acrónimo**	**Control**	**Tumor**	**Control**	**Tumor**
Cromófobo de riñón	KICH	25	66	25	66
Carcinoma escamoso de cuello y cabeza	HNSC	39	472	83	849
Carcinoma esofágico	ESCA	13	184	13	186
Carcinoma de riñón de célula papilar	KIRP	32	240	60	375
Carcinoma hepático	LIHC	50	268	99	460
Carcinoma de riñón de célula clara	KIRC	72	981	142	1.048
Adenocarcinoma de pulmón	LUAD	58	513	92	923
Carcinoma de tiroides	THAD	59	498	118	910
Adenocarcinoma de próstata	PRAD	41	333	102	635
Carcinoma urotelial de vejiga	BLCA	19	414	36	495
Carcinoma invasivo de mama	BRCA	113	1.134	207	2
Carcinoma escamoso de pulmón	LUSC	50	490	90	825
Adenocarcinoma de estómago	STAD	35	415	74	683
Cholangiocarcinoma	CHOL	9	36	9	36
Carcinoma endometrial de útero	UCEC	39	169	62	1.035
Total Muestras		**Control**	**Tumor**	**Control**	**Tumor**
18.605		654	6.213	1.212	10.526

Tabla 2. Conjunto de muestras seleccionas procedentes del Atlas genómico del cáncer (TCGA) utilizadas para llevar a cabo el presente estudio.

3.4. Rutas génicas relacionadas con cáncer.

Para poder estudiar las interacciones entre microARNs y genes relacionadas con el fenotipo tumoral, se recopilaron todos los genes involucrados en alguna de las siete rutas relacionadas con el crecimiento y la progresión tumoral mencionadas en el apartado de introducción y que son características de la naturaleza del cáncer (Hanahan and Weinberg 2011) y que se comentan brevemente a continuación:

1. **Ciclo celular**. Esta ruta esta constituida por diversas proteínas que controlan el periodo funcional de una célula hasta su división en dos células hijas. Según KEGG (Kanehisa, Sato et al. 2016) y Reactome (Fabregat, Sidiropoulos et al. 2016), proceden de un total de 582 genes.

2. **Ruta de respuesta al daño en el ADN**. Los genes que forman parte de esta ruta, codifican proteínas cuya función es detectar la presencia de lesiones en el ADN e intentar repararlas. Los 129 genes que definen esta ruta, se han obtenido a partir de la base de datos DDRProt (Andres-Leon, Cases et al. 2016), procedentes del trabajo de Aida Arcas et al (Arcas, Fernandez-Capetillo et al. 2014).
3. **Elongación de telómeros**. La renovación de los telómeros es un evento propio de células tumorales y se encuentra ligado a la sobreexpresión de diferentes genes, principalmente los correspondientes a telomerasas/polimerasas. El total de los 60 genes de esta ruta proceden de Reactome.
4. **Replicación del ADN**. Este acontecimiento tiene lugar de forma controlada en las células normales, Sin embargo, en el caso de las células tumorales, la tasa de replicación es tan elevada que la expresión de los genes implicados en este proceso suele aparecer profundamente alterados (Albertella, Lau et al. 2005). En concreto se caracterizaron 105 genes provenientes de Reactome.
5. **Senescencia**. Se trata de un proceso no proliferativo e irreversible que puede experimentar una célula. Asimismo, suele desencadenarse asociado al acortamiento de telómeros y evita la división celular. De acuerdo con Reactome, un total de 159 genes constituyen parte de esta ruta.
6. **Apoptosis**. Es el mecanismo de muerte celular programada más utilizada por las células sanas. Se activa con la finalidad de eludir la división en células que contienen errores en su material genético. En este caso, un total de 212 genes se obtuvieron desde KEGG y Reactome.
7. **Necrosis**. Junto con la apoptosis, la necrosis es un proceso encargado de la destrucción de células sanas en el momento en que determinadas señales de emergencia se producen. En esta ruta 21 genes derivados de Reactome han sido estudiados.

El conjunto total de genes relacionados con las distintas rutas génicas asociadas con cáncer se muestran en la Tabla Suplementaria 1 del Anexo I.

3.5. Análisis de las muestras procedentes de los tumores del TCGA.

Una vez que se hubo creado miRGate, una base de datos para obtener interacciones fiables entre miARN y ARNm, y miARma-Seq, una herramienta para identificar microARNs y genes diferencialmente expresados, se necesitaba procesar un elevado número de muestras de transcriptómica para obtener aquellos genes y miARNs implicados en el fenotipo

tumoral de un conjunto amplio de tumores y de esta forma, poder estudiar su red de regulación. A continuación se explican los pormenores del análisis de las muestras procesadas de los 15 tipos tumorales diferentes.

3.5.1. Análisis de las muestras de ARNm.

Un total de 6.867 muestras de RNASeq "*paired-end*", se procesaron usando miARma-Seq y siguiendo el protocolo elaborado por Anders et al, mayoritariamente empleado en este tipo de datos (Anders, McCarthy et al. 2013). Las muestras se procesaron con el objetivo de eliminar todas aquellas lecturas cuya calidad media fuera inferior a 20, las de alto contenido en nucleótidos sin identificar "N" y finalmente, seleccionando sólo aquellas secuencias en las que ambos pares estuvieran alineados con respecto al mismo fragmento genómico. Asimismo, el genoma humano usado como referencia, proviene de la versión 37 proporcionada por el consorcio de genomas de referencia (GRC) y descargado desde la versión 74 de EnsEMBL (Yates, Akanni et al. 2016).

Al concluir este proceso, se obtuvo el número de secuencias alineadas por cada uno de los genes expresados en la muestra. Dicho número, es proporcional al nivel de expresión del gen en cuestión. De tal forma que, la comparación del nivel de expresión de un gen (y por lo tanto del número total de secuencias alineadas) entre el grupo de muestras tumorales y muestras sanas, nos indica si debido al fenotipo tumoral, el número de secuencias alineadas y asociadas a este gen, se ha visto estadísticamente alterado entre ambos grupos de pacientes.

3.5.2. Análisis de las muestras de miARNs.

De forma análoga, 11.738 muestras de expresión de microARNs (miRNA-Seq o *small* RNA-Seq), se analizaron con miARma-Seq. En este caso, al tratarse de muestras secuenciadas mediante una técnica más sencilla (*"single-end"*), se excluyeron todas aquellas lecturas cuya calidad media fuera inferior a 20 o, por su alto contenido en nucleótidos "N". Igualmente, el genoma de referencia empleado procede de la versión 37 del consorcio de genomas de referencia y corresponde a la versión 20 de miRBase (Kozomara and Griffiths-Jones 2014). Además, como sucede en el análisis de la expresión diferencial de genes, el dato final, corresponde al valor de lecturas alineadas por cada miARN maduro en cada una de las muestras procesadas. El posterior análisis estadístico nos permite calcular aquellos microARNs que experimentan un cambio significativo de expresión entre los diferentes grupos de muestras.

3.5.3. Análisis estadístico.

Con el objetivo de determinar los miARNs o genes que muestran una variación estadísticamente significativa en sus niveles de expresión para un tipo de tumor específico, se examinó para cada tumor individual, el nivel de expresión (entendido como el número de lecturas alineadas) de cada gen/miARN procedente de la comparación entre el conjunto de muestras sanas y muestras tumorales. Para ello, se ha utilizado el módulo de expresión diferencial integrado en miARma y en concreto, el software de análisis diferencial edgeR (Robinson, McCarthy et al. 2010). En este proceso y según el trabajo de Anders et al. (Anders, McCarthy et al. 2013), los elementos de baja expresión, es decir aquellos con una sola lectura por cada millón de lecturas alineadas, fueron eliminados. Posteriormente, el resto de secuencias restantes fueron normalizadas mediante el algoritmo TMM (del inglés *trimmed mean of M-values)* propuesto por Robison et al. (Robinson and Oshlack 2010), con el objetivo de obtener valores comparables de lecturas alineadas independientemente de la profundidad de la secuenciación de cada muestra.

A continuación, se realizó un test estadístico basado en una distribución binomial negativa (Robinson and Smyth 2008) para calcular las diferencias entre ambos grupos de muestras, así como el cambio de expresión experimentado por los genes o miARNs (representado en logaritmo en base 2 ó logFC) junto con el valor de probabilidad corregido en base al algoritmo propuesto por Benjamini y Hochberg (Hochberg and Benjamini 1990) denominado falso ratio de descubrimiento o FDR (del ingles *False Discovery Rate)*. Por último, aquellos miARNs o genes con un valor de probabilidad FDR inferior a 0.05, fueron considerados estadísticamente relevantes y por lo tanto expresados diferencialmente. De estos, se seleccionaron aquellos genes o miARNs que hubieran experimentado un cambio de expresión relevante, mediante la elección de aquellos con un valor de cambio de expresión absoluto logFC, iguales o superiores a 1.

3.6. Estudio funcional de las rutas relacionadas con cáncer.

En el presente trabajo, se determinó la relevancia de la expresión diferencial del conjunto de genes, mediante un estudio funcional. Dado que los genes del presente estudio fueron seleccionados en base a siete rutas esenciales en la biología del cáncer y procedentes de tres fuentes diferentes, estas difieren de las vías ofrecidas por los programas especializados en estudios de enriquecimiento. Consecuentemente, incorporamos un análisis híper-geométrico basado en un enfoque ampliamente extendido para el análisis de rutas en muestras de cáncer (Pau Creixell 2015). De esta forma, se calculó la razón de

probabilidades (OR, del inglés *odds ratio*) mediante el uso del test exacto de Fisher en función de una tabla de contingencia de 2x2. Este test nos posibilita verificar si la comparación entre el número de genes desregulados y no alterados, que forman parte de una determinada ruta, difiere del valor obtenido entre número total de genes diferencialmente expresados en el conjunto del experimento, frente al total de genes. Este test además proporciona un valor estadístico ajustado para aceptar o rechazar la hipótesis nula (H_0) que implicaría que los genes diferencialmente expresados se distribuyen de forma independientemente de la ruta a la que pertenecen. Tras descartar la hipótesis nula con valores de P. ajustado <=0.05 y dependiendo del OR obtenido, podemos comprobar el índice de regulación a la que cada ruta está sometida. De tal forma que un OD con un valor de 3, implicaría que el número de genes diferencialmente expresados (ya sean sobre-expresados o reprimidos) de esa ruta, es tres veces superior al valor esperable en base al resto de genes desregulados del experimento. En tal caso se concluiría que se trata de una ruta alterada de forma importante y relacionada con el fenotipo a estudiar. Sin embargo, valores OR inferiores a 1, como 0.5, nos indican que el número de genes desregulados en esa ruta es estadísticamente inferior (exactamente la mitad) al número esperado en función del número total de genes diferencialmente expresados. En este ejemplo, se concluiría que los genes que constituyen esa ruta, presentan de forma no esperable, unos valores de expresión equivalente al de las muestras control.

Asimismo como control, se escogió de forma aleatoria del total de genes desregulados en cada estudio, un número de genes idéntico al de cada una de las siete rutas, lo que permite comprobar tanto si los valores de probabilidad como la razón de probabilidades son verdaderamente indicativos y específicos en el conjunto de rutas seleccionadas. Los valores de OD, así como los valores de probabilidad para cada ruta y tumor, se muestran en la Tabla Suplementaria 2 del Anexo I.

3.7. Análisis integrado de las muestras procedentes de los tumores del TCGA.

3.7.1 Análisis integrado del conjunto de tumores.

Una vez analizadas las muestras de RNA-Seq y de miRNA-Seq para cada uno de los 15 tipos de tumor, se obtuvo un listado de genes y miARNs alterados, que fueron estudiados en un marco Pan-cáncer. Como la mayoría de los genes y microARNs relacionados con el desarrollo y la progresión tumoral, aparecen fuertemente desregulados en muestras tumorales (Lee and Young 2013), se seleccionaron todos aquellos genes y miARNs

diferencialmente expresados con un valor de FDR ≤ 0.05 y un valor de log2FC ≥ 1. Posteriormente, para investigar la implicación de estos genes y miARNs en el cáncer, se seleccionaron solo aquellos que aparecían diferencialmente expresados de forma conjunta en un alto número de tipos tumorales. Con el fin de identificar este número de corte necesario, estudiamos la listas de genes y miARNs obtenidos a diferentes umbrales (entre 5 y 10 tipos tumorales) y calculamos el enriquecimiento en elementos previamente conocidos por estar relacionados con el cáncer mediante el uso de un test de Fisher de cola derecha.

En el caso de los genes, seleccionamos dos fuentes de datos que proporcionan una lista de genes relacionados con el cáncer, como es el censo de genes del cáncer de COSMIC (Forbes, Beare et al. 2015) y la red de genes del cáncer, NCG 5.0 (An, Dall'Olio et al. 2016), la cual recoge 1571 genes del cáncer (518 genes conocidos y 1053 genes candidatos) de un total de 175 estudios publicados. Una vez obtenidos, se estudió el enriquecimiento en los genes conocidos del cáncer que se obtenían a los distintos umbrales.

Para el estudio de los miARNs el enfoque fue muy similar. A través de la base de datos OncomirDB (Wang, Gu et al. 2014), se obtuvo el conjunto de miARNs relacionados con el cáncer y posteriormente, se estudió el enriquecimiento que se obtenía a distintos umbrales.

Como resultado en ambos casos, los valores más altos de la razón de probabilidades u *Odd ratio* estadísticamente significativos (FDR<0.05), se obtuvieron cuando el nivel de corte se fijaba en 8 tipos tumorales (Tabla 3a y 3b). De forma que, del conjunto de genes y miARNs diferencialmente expresados en los 15 tipos de tumores, se seleccionaron sólo aquellos que aparecieran desregulados de forma conjunta en al menos 8 tipos de tumores distintos.

Por otra parte, para disminuir el número de falsos positivos obtenidos a partir de los algoritmos de predicción, y teniendo en cuenta que los microARNs disminuyen la expresión de los genes, el proceso de predicción de interacciones empleó perfiles de expresión invertida entre genes y miARNs, tal y como recomiendan otros autores (Vishnubalaji, Hamam et al. 2015). Además, solo se consideraron aquellas interacciones obtenidas por más de un método de predicción en la misma coordenada genómica (concordancia genómica >= 2) y cuando fuera posible, por isoformas principales según GENCODE, sub-proyecto de la enciclopedia de elementos del ADN encargado de la anotación de genes e isoformas en humano y ratón (Pei, Sisu et al. 2012).

Tabla 3. Enriquecimiento relacionado con el cáncer de genes y microARNs, cuantificado a diferentes umbrales (número de tipos tumorales.)

3a. Enriquecimiento relacionado con el cáncer, de los genes obtenidos a distintos umbrales. Cuando el umbral se fija en 8, los genes diferencialmente expresados obtenidos en al menos 8 tipos tumorales en conjunto, obtienen el mejor valor de *odd ratio.*

Umbral	Número de genes	Genes del cáncer	FDR	Odd Ratio	Fuente
5	226	19	8.16E-03	3.13	COSMIC
5	226	37	1.26E-03	2.45	NCG
6	191	15	9.37E-02	2.89	COSMIC
6	191	32	2.83E-03	2.5	NCG
7	169	14	7.49E-02	3.1	COSMIC
7	169	29	3.96E-03	2.6	NCG
8	147	14	**1.54E-02**	**3.56**	COSMIC
8	147	27	**1.87E-03**	**2.8**	NCG
9	131	12	6.03E-02	3.4	COSMIC
9	131	21	1.01E-01	2.36	NCG
10	107	10	1.24E-01	3.27	COSMIC
10	107	16	6.70E-01	2.17	NCG

3b. Enriquecimiento relacionado con el cáncer de los miARNs obtenidos a distintos umbrales. Cuando el umbral se fija en 8, los miARNs diferencialmente expresados obtenidos en al menos 8 tipos tumorales en conjunto, obtienen el mejor valor de *odd ratio.*

Umbral	Número de miARNs	miARNs en cáncer	FDR	Odd Ratio	Fuente
5	268	100	1.04E-28	5.91	OncomirDB
6	190	79	3.00E-25	6.55	OncomirDB
7	132	56	1.09E-17	6.29	OncomirDB
8	95	47	**5.25E-18**	**8.13**	OncomirDB
9	61	30	3.21E-11	7.67	OncomirDB
10	36	18	7.33E-07	7.68	OncomirDB

Asimismo, con el fin de obtener interacciones miARN-ARNm de alta especificidad, se aplicó el siguiente enfoque estadístico: para cada socio, ya sea un miARN o un ARNm que forma parte de una interacción, se obtuvieron todos los microARNs que podrían regular al gen de estudio y de forma contraria, todos aquellos genes, regulados por el microARN de interés, que aparecieran en un mínimo de 8 tipos tumorales distintos. Mediante la utilización de un test exacto de Fisher con una tabla de contingencia de 2x2, obtendríamos como pares estadísticamente relevantes, aquellas formados por genes regulados por un número significativamente bajo de microARNs y a su vez, por miARNs que regulan un ínfimo número de genes. Estas interacciones formadas por genes y miARNs con pocos

interactores secundarios, son esenciales para el desarrollo de nuevas terapias capaces de disminuir efectos no deseados. Tabla Suplementaria 7 del anexo I.

3.7.2 Análisis integrado de tumores procedentes de un mismo origen.

Además de estudiar aquellas interacciones que aparecen en un alto número de tipos tumorales distintos, los genes y miARNs diferencialmente expresados en cada tipo de tumor individual con un FDR <= 0.05 y una variación significativa en la expresión (log2FC absoluto >= 1) fueron empleados para computar todas las posibles interacciones procedentes de miRGate (Andres-Leon, Gonzalez Pena et al. 2015). A continuación, se seleccionaron aquellas asociaciones resultantes de más de un algoritmo en el misma posición genómica del 3'-UTR (concordancia genómica > = 2). Por último, una vez obtenido este conjunto de interacciones de confianza, se compararon entre cada uno de los 15 tipos tumorales para obtener una lista de interacciones miARN-ARNm exclusivas para cada tipo de tumor. Esto permitió hacer énfasis en aquellos tipos tumorales procedentes de un mismo origen, como es el caso de pulmón (con muestras derivadas de adenocarcinoma de pulmón y tumor escamoso) o riñón (con muestras de tipo cromófobo, papilar y de célula clara).

3.7.3 Análisis del efecto de la metilación y de la alteración en el número de copias en las interacciones identificadas.

La asociación entre la expresión del gen y del microARN procedente de las interacciones obtenidas, fue analizada mediante el empleo de una regresión lineal multifactorial incluyendo los valores de expresión del gen (logaritmo de las "secuencias por kilo-base de transcrito por cada millón de secuencias alineadas" o RPKM, del inglés, *Reads Per Kilobase of transcript per Million mapped reads*) como variable respuesta y los valores de expresión del miARN (logaritmo del valor de expresión medido en RPKM) como predictor. Para corregir el efecto que sobre la expresión del gen pudiera tener la metilación diferencial de este o la alteración en el número de copias (CNAs), ambas variables fueron incluidas en el modelo de regresión.

Los datos de CNAs, analizados con Gistic2 (del inglés *Genomic Identification of Significant Targets in Cancer*) (Mermel, Schumacher et al. 2011) se descargaron del portal de datos del TCGA cBioportal (Cerami, Gao et al. 2012, Gao, Aksoy et al. 2013), y fueron categorizados en cinco niveles diferentes: deleción homocigótica (-2), deleción hemicigótica (-1), sin cambio (0), ganancia (1) y amplificación de alto nivel (2).

En el caso de los datos de metilación del ADN, éstos se obtuvieron de Wanderer (Diez-Villanueva, Mallona et al. 2015), una herramienta web que ofrece los valores de metilación en el ADN de las muestras procedentes del TCGA para los genes humanos. Sin embargo, debido a que Wanderer carecía de los datos de la metilación de los tumores de cholangiocarcinoma (CHOL), esta información se obtuvo de cBioportal. Posteriormente, estos datos fueron transformados de los valores β proporcionados, a valores M para reducir la heterocedasticidad (para contrarrestar la variabilidad entre muestras) antes de ser incluidos en los consiguientes análisis (Du, Zhang et al. 2010).

De esta forma, para cada interacción ARNm-miARN, se realizó un análisis de regresión lineal individual para cada uno de los 15 tipos de canceres del estudio y posteriormente un análisis adicional conjunto, que incluía todas las muestras tumorales. En este último tipo de análisis, con el objetivo de tener en cuenta las diferencias entre la expresión génica y la expresión del miARN entre los distintos tipos tumorales, el tipo de cáncer y la interacción estadística entre este y la expresión del microARN, fueron incluidos como factores adicionales. En todos los análisis, la correlación de expresión entre el gen y el miARN, fue calculada como la raíz cuadrada de la fracción de varianza de la expresión del gen explicada por la expresión del miARN, después de corregir el efecto de los otros predictores (tipo de cáncer y/o CNA y metilación) mediante un análisis de varianza (o ANOVA). El sentido de la correlación, ya sea positivo o negativo, así como el estadístico t y el valor de probabilidad resultante, se obtuvieron directamente de los coeficientes de los modelos lineales. Los P-valores, posteriormente se ajustaron por el método FDR (Benjamini Y 1995). En algunos casos, debido a la ausencia de expresión del gen o del microARN en algún tipo tumoral, los modelos de regresión lineales no pudieron ser calculados, y en ese caso se consideró que no existía correlación entre la expresión del gen y la del microARN (Figura 23 y 28a).

3.7.4 Análisis de supervivencia.

El efecto de la expresión del gen y del miARN regulador en la supervivencia del paciente, se analizó mediante el empleo de los modelos de riesgos proporcionales de Cox para cada uno de los tipos tumorales de forma individual. Todos los modelos calculados incluyeron el estadío tumoral como cofactor, codificado en dos categorías: buen pronóstico (estadíos I o II) y mal pronóstico (estadíos III, IV y X). En los cánceres de próstata, dada la ausencia del estadío tumoral, se empleó el grado tumoral, clasificado según la escala de Gleason. De

esta forma los valores 6 y 7 fueron categorizados como buen pronóstico y los superiores a 7, como mal pronóstico.

En estos análisis, los valores de expresión de los genes y de los microARNs se introdujeron como variables cuantitativas (RPKMs transformados logarítmicamente). El análisis multi-variable de los riesgos proporcionales de Cox, nos proporciona una estimación del efecto de la expresión del gen / miARN en la supervivencia, independientemente del efecto del estadio tumoral (o grado tumoral, en el caso del cáncer de próstata). Los valores de probabilidad obtenidos, fueron posteriormente ajustados por el método del falso ratio de descubrimiento o FDR (Benjamini Y 1995). Figura 27 y 28b.

4. RESULTADOS

En esta sección, se explicará en profundidad los resultados obtenidos para cada uno de los objetivos específicos propuestos en el apartado 2 de esta tesis doctoral. Para ello se expondrán los resultados de nuestra base de datos denominada miRGate, orientada al estudio de las interacciones que se establecen entre miARNs y genes, haciendo especial énfasis en la fiabilidad mostrada por esta base de datos. En segundo lugar, se mostrará la alta correlación de los datos generados mediante el uso de la herramienta miARma-Seq, diseñada para el análisis de muestras de expresión, procedentes de técnicas de secuenciación masiva como son: RNASeq y miRNASeq. A continuación, se mostrarán los resultados alcanzados tras aplicar conjuntamente ambas herramientas en un amplio grupo de muestras de diferentes tipos canceres de pacientes procedentes de El Atlas Genómico del Cáncer (TCGA). De esta forma, la sección de resultados constará de tres apartados cuya relación con los objetivos propuestos aparecen detallados en la Tabla 4.

Objetivo	Herramientas desarrolladas	Artículo
Metodología capaz de predecir interacciones entre miARNs y ARNm fiables.	miRGate	Andrés-León, 2015 Andrés-León, 2017b
Herramienta computacional para el estudio de muestras de expresión de genes y miARNs procedentes de muestras tumorales y control de pacientes.	miARma-Seq	Andrés-León, 2016
Obtener un conjunto de genes y microARNs diferencialmente expresados entre muestras sanas y tumorales de pacientes.		
Estudio global de genes y microARNs constitutivamente desregulados en diferentes conjuntos de tumores y las posibles interacciones de regulación entre ellos.		Andrés-León, 2017a
Corrección de factores de confusión como la metilación y la alteración del número de copias, en la expresión génica.		
Estudio de la supervivencia de los pacientes. En función de las asociaciones obtenidas los diferentes conjuntos de tumores.		

Tabla 4. Equivalencia entre los objetivos específicos propuestos, con las herramientas creadas y los trabajos publicados.

4.1. miRGate: base de datos que almacena interacciones miARN-ARNm fiables para humano, rata y ratón.

miRGate almacena predicciones de interacciones establecidas entre miARNs y ARNm. Este repositorio se caracteriza por el uso de cinco algoritmos distintos de predicción, ampliamente usados en el área de la regulación por miARNs, como son: miRanda (Enright, John et al. 2003), Pita (Kertesz, Iovino et al. 2007), RNAHybrid (Kruger and Rehmsmeier 2006), TargetScan (Friedman, Farh et al. 2009) y microTar (Thadani and Tammi 2006) respaldado, sólo en el caso de Pita, por la validación en el laboratorio de un conjunto pequeño de sus predicciones más destacadas. Sin embargo, no existe una comprobación de la fiabilidad del total de sus predicciones, en base a un conjunto de asociaciones previamente confirmadas experimentalmente.

Por otro lado, resulta importante destacar que las predicciones recopiladas en miRGate, aunque hayan sido computadas por los mismos algoritmos, proceden de un conjunto de secuencias distinto al empleado por cada programa de predicción (Más detalles en la Tabla 1). Por lo tanto, estas nuevas predicciones almacenadas en miRGate también fueron evaluadas frente al mismo conjunto de interacciones confirmadas experimentalmente.

4.1.1. Las predicciones de miRGate están enriquecidas en interacciones validadas experimentalmente.

En el presente trabajo, para cuantificar la fiabilidad de las interacciones, fue necesario obtener un conjunto amplio de asociaciones comprobadas experimentalmente. Estas se obtuvieron de los cuatro repositorios de uso más extendido, como son: Tarbase (Vergoulis, Vlachos et al. 2012), miRTarBase (Hsu, Tseng et al. 2014), miRecords (Xiao, Zuo et al. 2009) y OncomirDB (Wang, Gu et al. 2014). El número total de las interacciones validadas entre genes y miARNs humanos almacenados en estos repositorios es de 79.046, entre ellas, eliminando todas las asociaciones repetidas, obtenemos un total de 40.991 (52%) ARNm–miARN diferentes. (Figura 11).

Procedente de miRGate se recopilaron los valores estandarizados obtenidos en cada predicción, y se representaron mediante una curva tipo ROC (del inglés *Receiver Operating Characteristic*). Las curvas ROC, se definen como una representación gráfica de la sensibilidad frente a la especificidad para un sistema clasificador binario (predicción correctamente validada o no) según se modifica el umbral de discriminación. Por consiguiente, es una forma de evaluar el ratio de verdaderos positivos (VPR = Razón de Verdaderos Positivos) frente a la razón de falsos positivos (FPR = Razón de Falsos

Positivos). Asimismo, estas curvas nos proporcionan un valor que permite representar la probabilidad de que una predicción sea correcta y por consiguiente confirmada experimentalmente, frente a una predicción incorrecta, denominado el Área Bajo la Curva (AUC, del inglés *Area Under the Curve*). De esta forma, un AUC con valor de 1, implica que el total de las predicciones evaluadas son correctas, mientras que valores de 0.5, conllevan un cierto grado de aleatoriedad presente en los resultados. En la Figura 12 se muestran los valores de AUC obtenidos por el conjunto de predicciones almacenadas en miRGate, frente a los valores generados por cada método de forma independiente. Tras este análisis, el área bajo la curva obtenido a partir de los datos de miRGate es de 0.704, mientras que el de microTar es de 0.528, RNAHybrid 0.609, miRanda 0.632, Targetscan 0.638 y finalmente, Pita 0.548. Las diferencias entre los métodos se incrementan considerablemente, si se comparan los valores de AUC obtenidos por las predicciones almacenadas en miRGate, frente a las ofrecidas por cada método y disponibles en la red.

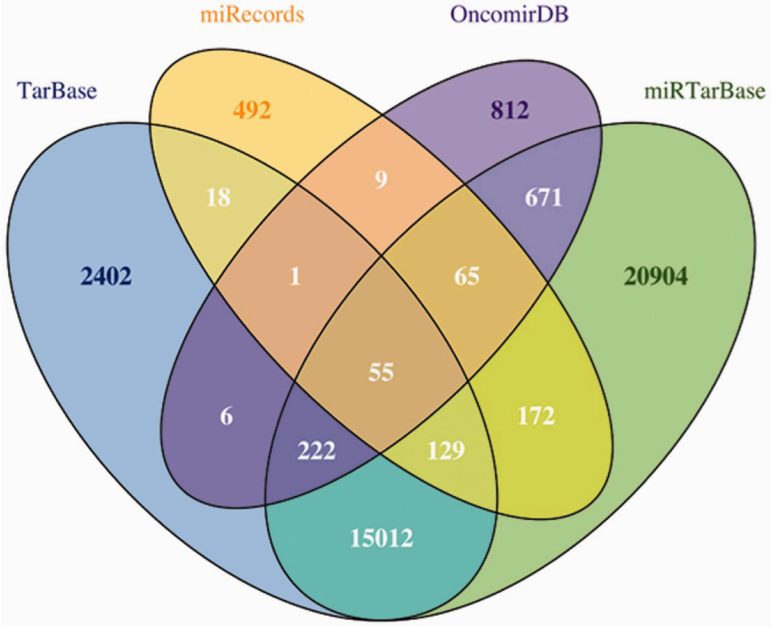


Figura 11. Diagrama de coincidencia entre las predicciones confirmadas experimentalmente. El diagrama de Venn adjunto representa la bajo coincidencia entre las cuatro bases de datos que compilan interacciones validadas en laboratorio, como son Tarbase, miRTarBase, miRecords y OncomirDB **.**

Como se puede apreciar en la Figura 13a, mientras el valor de AUC de miRanda es de 0.599, Targetscan de 0.56 y Pita de 0.63, el de miRGate es de 0.704.

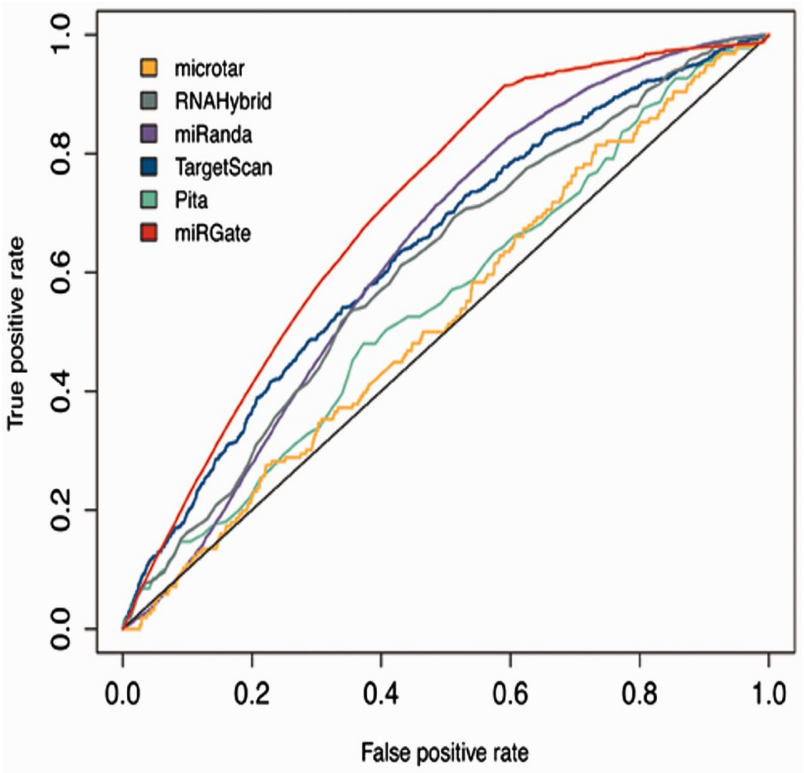


Figura 12. Comparación de la curva ROC obtenida con miRGate frente a sus métodos.
Esta curva presenta el valor de fiabilidad de las predicciones almacenadas en miRGate en conjunto, frente a las predicciones de métodos usados por separado, cuando estas se enfrentan a datos validados en laboratorio. El área bajo la curva de los distintos métodos muestra los valores siguientes: microTar: 0.538, RNAHybrid: 0.609, miRanda: 0.632, Targetscan: 0.638, Pita: 0.548, y por último miRGate: 0.704

Otra característica relevante a exponer, es que los valores de AUC obtenidos, dependen de la fiabilidad de las bases de datos que proporcionan las interacciones validadas, y a su vez, de los métodos empleados por éstas, para almacenar las predicciones confirmadas. Como ya se mostró en la Figura 11, la baja coincidencia entre las distintas bases de datos que proporcionan datos validados experimentalmente es significativa, a pesar, incluso, de obtener la información de la misma fuente, como es Pubmed. Por ese motivo, a continuación se elaboró una clasificación de los repositorios en función de un criterio de fiabilidad, y de esta forma, OncomirDB fue considerado un repositorio con elevada fiabilidad ya que las interacciones proporcionadas habían sido seleccionadas manualmente por investigadores expertos. Sin embargo, miRecords, fue categorizada con una fiabilidad media, debido a que sus predicciones aunque habían sido seleccionadas por expertos, éstas procedían de artículos científicos procesados automáticamente. Por último, TarBase y miRTarBase fueron clasificadas como bases de datos de menor fiabilidad debido a que las interacciones almacenadas habían sido obtenidas a través del uso de técnicas automatizadas basadas en la minería de datos. En la Figura 13b se muestra como el AUC del conjunto de predicciones de miRGate aumenta hasta 0.769, si sus datos son equiparados a un repositorio de alta fiabilidad. Del mismo modo, frente a miRecords, de fiabilidad media, el AUC obtenido por la información almacenada en miRGate es de 0.727 y en el caso de utilizar repositorios de menor confianza en la evaluación, el AUC disminuya a 0.699.

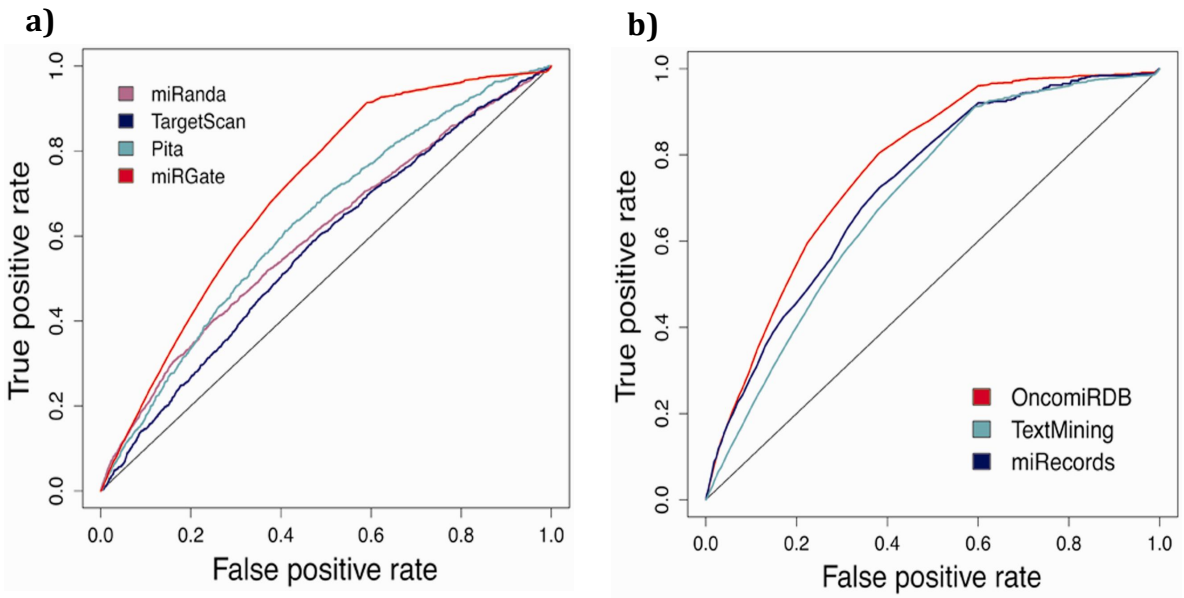


Figura 13. Curvas ROC generadas a partir de las predicciones de miRGate en comparación con otras fuentes. a| miRGate proporciona un valor de fiabilidad superior sobre las predicciones, en comparación con las proporcionadas por otros algoritmos. El AUC de miRGate alcanza un valor de 0.704, 0.599 para miRanda, 0.56 para Targetscan y 0.63 para Pita. b| El valor de fiabilidad varía en función de la confianza de los repositorios empleados. El AUC obtenido frente a OncomirDB, una base de datos de elevada fiabilidad es de 0.769, miRecords, de fiabilidad media, presenta un AUC de 0.727, que disminuye a 0.699 si la fiabilidad de los repositorios es baja.

Por último, resulta importante destacar que, en diversas colaboraciones establecidas con otros investigadores, miRGate proporcionó predicciones que posteriormente fueron validadas con éxito en diferentes laboratorios, tal y como se ha mencionado en el apartado 3.1.4. Entre estas colaboraciones, mostró gran relevancia la contribución aportada al estudio del cáncer de mama triple negativo, realizado por el grupo del Dr. Javier Benítez en el Centro Nacional de Investigaciones Oncológicas (CNIO) en Madrid. Este proyecto de investigación, se basaba en el uso de microarrays pertenecientes a 11 pacientes sanos y 122 enfermos, en el que se detectó la expresión diferencial de 105 microARNs. Entre las predicciones proporcionadas por miRGate y en función de su capacidad de regular BRCA1, miRGate destacó tres miARNs por tener altos valores centralizados (Figura 14a): miR-498 (con dos posibles sitios de interacción), miR-187-5p (con un único sólo sitio de unión) y miR-146a-5p (con dos sitios de unión previamente validados por otros autores). De tal forma, que a través de un ensayo de actividad de luciferasa, se confirmó que en presencia de los tres microARNs propuestos por miRGate, la expresión de BRCA1 disminuía entre un 40 y un 60% (Figura 14b).

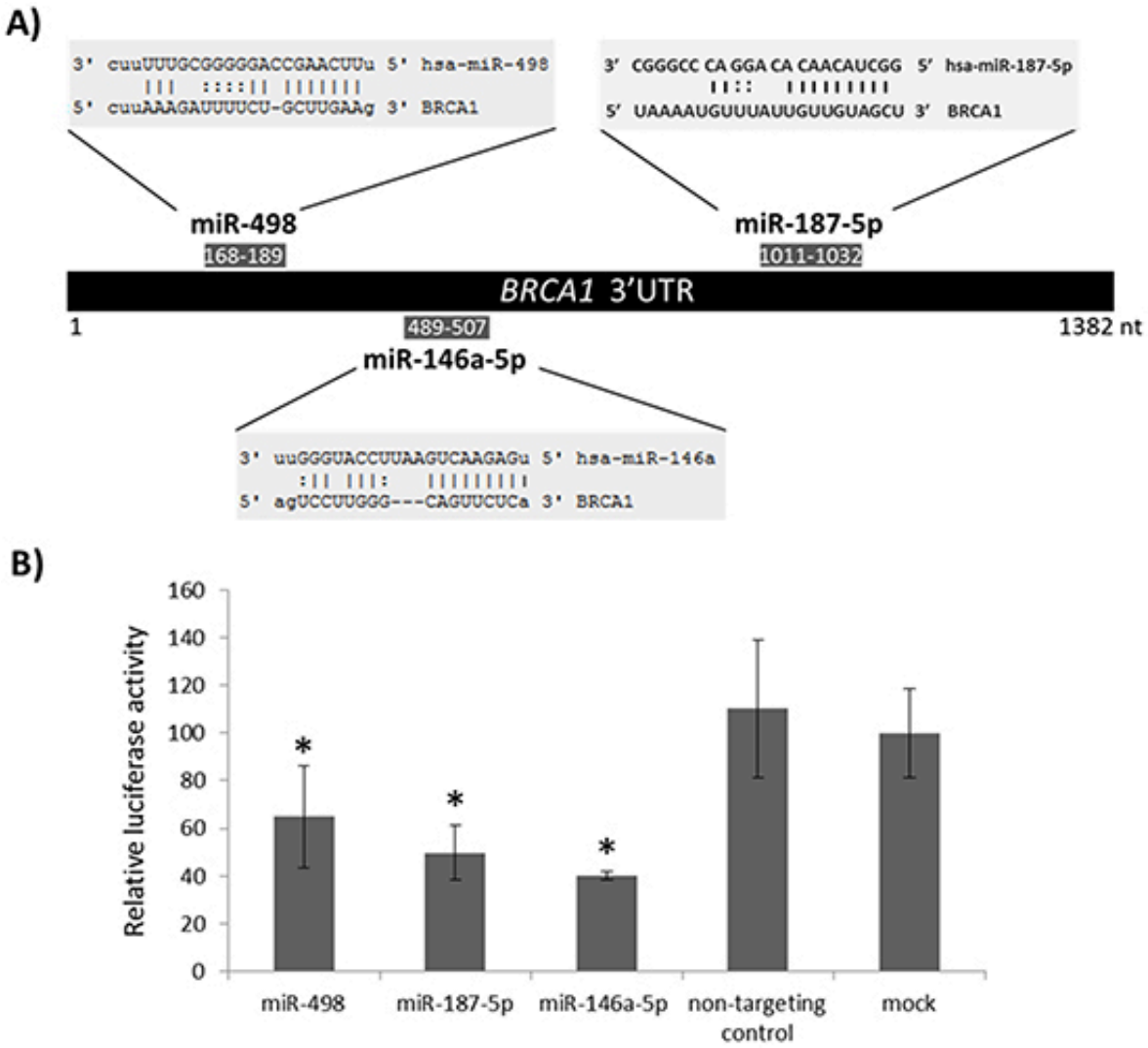


Figura 14. Validación experimental de las predicciones generadas por miRGate. a| Se representa las tres predicciones con mejor valor centralizado según miRGate asociadas a BRCA1. Además de la secuencia del miARN y de la 3'UTR de BRCA1, se indican las coordenadas genómicas donde se establecen las uniones. b| Valores de expresión según la actividad luciferasa de los tres precursores de miARNs seleccionados. En la comparación se incluye una muestra control de células 293T y una muestra con transfección de luciferasa pero sin precursor (mock). * $p < 0.05$. Adaptada de Matamala et al *Oncotarget* (2016).

Por otra parte, debido a que miRGate es una potente herramienta que permite generar predicciones incluso a partir de miARNs víricos, la Dra. Giuglia di Falco de la Universidad de Siena, solicitó nuestra participación en un estudio basado en la investigación de linfomas asociados a infecciones producidas por el virus Epstein-Barr. Previamente, otros autores han descrito que el 90% de los linfomas de Burkitt endémicos y el 20% de los esporádicos, están relacionados con la infección causada por el virus Epstein-Barr. En esta colaboración, se realizó un estudio de expresión diferencial de muestras de linfoma positivas para este virus, frente a muestras de linfoma sin infectar, que dio lugar a un total

de 8 microARNs humanos y 11 miARNs pertenecientes al virus Epstein-Barr, que presentaban una expresión estadísticamente alterada.

Entre los genes dianas proporcionados por miRGate, la mayoría estaban involucrados en procesos de ciclo celular, diferenciación, proliferación celular y apoptosis. Concretamente, el miARN ebv-BART6-3p destacó por ser el único microARN que regulaba a diferentes genes, cuyas funciones estaban relacionadas con todas las rutas mencionadas anteriormente.

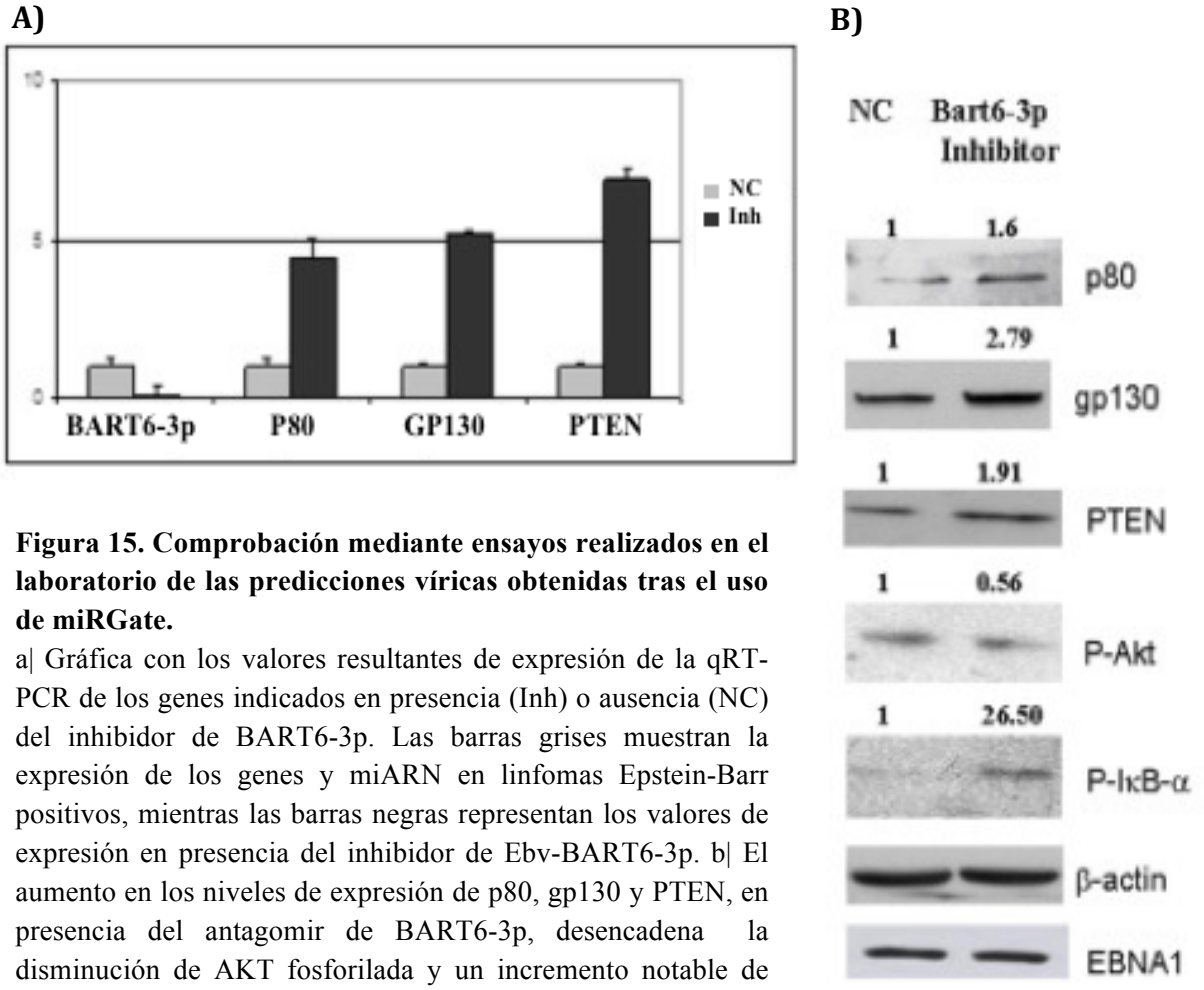


Figura 15. Comprobación mediante ensayos realizados en el laboratorio de las predicciones víricas obtenidas tras el uso de miRGate.

a| Gráfica con los valores resultantes de expresión de la qRT-PCR de los genes indicados en presencia (Inh) o ausencia (NC) del inhibidor de BART6-3p. Las barras grises muestran la expresión de los genes y miARN en linfomas Epstein-Barr positivos, mientras las barras negras representan los valores de expresión en presencia del inhibidor de Ebv-BART6-3p. b| El aumento en los niveles de expresión de p80, gp130 y PTEN, en presencia del antagomir de BART6-3p, desencadena la disminución de AKT fosforilada y un incremento notable de NFK-b fosforilada. Adaptada de Ambrosio et al. *Infect Agent Cancer*. (2014).

Además, según miRGate, este miARN regulaba entre otros a *PTEN* y a ambas cadenas del receptor de *IL-6* (p80 y gp130). A continuación, la validación de estas predicciones se llevó a cabo con un antagomir (AMO) específicamente diseñado para inhibir la función de BART6-3p. De tal forma que tras 24 horas de la transfección con el inhibidor del miARN, los niveles de expresión de éste disminuyeron considerablemente (Figura 15a). Asimismo,

la completa inhibición de la expresión de BART6-3p, desencadenó el aumento de los niveles de expresión tanto de *PTEN* como de *P80* y *GP130*, confirmando de esta manera la regulación por este miARN. Por otra parte, el incremento de expresión del receptor de *IL-6*, provocó la activación de la ruta NF-kappa B. Además, dado que PTEN regula negativamente la activación de Akt a través de su fosforilación, su activación debido a la inhibición de BART6-3p, dio lugar a una elevación en los niveles de Akt fosforilada y por consiguiente, a una disminución de su actividad.

4.1.2. Acceso fácil y versátil a miRGate desde la web.

Las predicciones generadas por miRGate (Tabla 5) se encuentran almacenadas en una base de datos relacional en MySQL. Con el fin de facilitar el acceso a esta información por parte de otros investigadores interesados de una forma rápida y sencilla, se ha desarrollado una página web. Esta página ha sido diseñada con el objetivo de proporcionar la mayor cantidad de información posible, precisando por el contrario, de datos mínimos por parte del usuario. La página web consta de diferentes e intuitivos pasos, donde es necesario seleccionar el tipo de organismo de estudio, junto con el identificador del gen y/o del miARN a analizar. Asimismo, la web acepta distintos identificadores, por ejemplo, para un gen soporta el uso del nombre según HGNC (del inglés *HUGO Gene Nomenclature Commitee)*, identificadores de EnsEMBL (tanto para genes cómo para transcritos) y nombres de sondas para microarrays de las compañías Affymetrix e Illumina. Sin embargo, en la búsqueda de miARNs, miRGate admite identificadores de miRBase, ya sean identificadores de pre-miARNs o miARNs maduros, así como nombres de sondas de microarrays de la empresa Agilent.

Por otra parte, la página web permite filtrar los resultados según distintos parámetros ofrecidos al usuario. Entre las alternativas posibles, se puede elegir el/los método/s de predicción (o repositorios de información experimental) a mostrar, el valor mínimo de concordancia genómica y el tipo de transcrito a mostrar (si codifica a una proteína, se trata de un pseudogen, usando isoformas principales según GENCODE de forma estándar), entre otros.

Método \ Organismo	Humano	Ratón	Rata
miRanda	34.838.559	16.164.311	1.372.897
Pita	773.112	313.113	52.281
RNAHybrid	36.832.689	10.390.354	536.248
microTar	6.049.837	1.750.058	3.348.100
Targetscan	7.270.936	5.186.036	417.501
TarBase	36.853	20.513	7
miRTarBase	39.118	9.314	307
miRecords	1.198	227	-
OncomirDB	2.368	1.917	-
miRGate	85.844.670	33.835.843	5.727.341

Tabla 5. Número total de predicciones obtenidas por miRGate clasificadas por organismo y algoritmo utilizado.

4.1.3. Acceso programático a las predicciones de miRGate.

En la actualidad, las técnicas de transcriptómica posibilitan la medición de lo niveles de expresión de una elevada cantidad de microARNs y genes, e incluso de sus variantes de expresión (transcritos) de forma simultánea. Sin embargo, el gran volumen de datos generados, dificulta considerablemente la consulta manual de interacciones entre miles de genes y/o miARNs a través de una página web, afectando principalmente a la velocidad de la consulta. Par eliminar estas limitaciones y conseguir predicciones de miles de genes y/o miARNs de forma eficaz, miRGate incluye un Interfaz de Acceso Programático (API) basado en la tecnología del lenguaje de marcas extensible o XML (del inglés *eXtensible Markup Language)*. De tal forma que, mediante un protocolo de transferencia de estado representacional (REST) se permite la ejecución de consultas a la base de datos de manera directa desde cualquier tipo de lenguaje de programación (Andres Leon, Gomez-Lopez et al. 2017). Además, la versión actual a parte de la obtención de interacciones, facilita la recuperación inmediata de información sobre miARNs y genes, como son sus coordenadas genómicas, sinónimos de identificadores y la obtención de secuencias completas o secuencias semilla en el caso de los microARNs. Por último, las especificaciones para el uso del conjunto de opciones ofrecidas, así como diversos ejemplos para realizar consultas en base al protocolo REST, y la documentación, se encuentran disponibles en la página web del API de miRGate: http://mirgate.bioinfo.cnio.es/API/api.html.

4.1.4 Estadísticas de uso de miRGate.

Una forma adicional de valorar los resultados que miRGate ofrece a los investigadores, puede medirse por el nivel del uso de la herramienta desde el día de su publicación. Para ello hemos obtenido el número de peticiones, ya sea desde la web o desde el API de consulta, que han sido proporcionadas desde la base de datos, eliminado todas aquellas realizadas por indexadores automáticos tipo Google. Desde Mayo de 2015 hasta Enero de 2017, miRGate ha recibido un total de 6.9 millones de peticiones (~370.000 peticiones/mes) tal y como se muestra en la Figura 16.

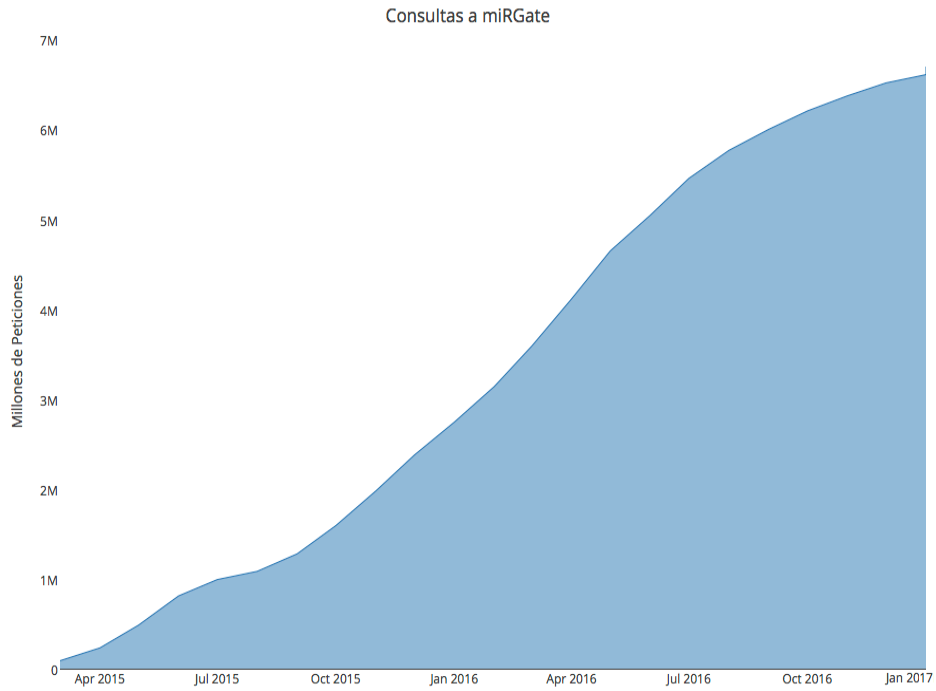


Figura 16. Estadísticas de uso de miRGate. El número de peticiones a miRGate desde su publicación, hasta Enero de 2017 es de cerca de 7 millones.

4.2. miARma-Seq, una herramienta efectiva para el análisis sistemático de microARNs, ARNm y ARNs circulares.

Para realizar parte de los objetivos expuestos, es necesario obtener el conjunto de genes y microARNs diferencialmente expresados en un elevado número de muestras tumorales. Debido a que el análisis pormenorizado y exhaustivo de este volumen de muestras es inviable, y que las muestras a analizar (RNASeq o miRNASeq) difieren en gran medida, se desarrolló una herramienta llamada miARma-Seq, capaz de realizar estos análisis de forma rápida y eficaz. Para comprobar la validez de los resultados obtenidos por miARma-Seq se evaluó cada uno de los módulos de análisis frente a un conjunto de datos confirmados

experimentalmente. Particularmente, dado el bajo número de resultados validados en este tipo de análisis, se calculó la correlación entre los valores de expresión resultantes tras el uso de miARma-Seq con los valores obtenidos por otros autores que, además de utilizar herramientas de análisis diferentes, comprobaron experimentalmente parte de los resultados.

4.2.1. Validación de datos de expresión de microARNs.

La capacidad de miARma-Seq para identificar, cuantificar y obtener miARNs diferencialmente expresados, se evaluó mediante el análisis de los datos procedentes de la expresión de microARNs en muestras de tumor de mama bajo condiciones de hipoxia (oxígeno 1%), conseguidos mediante el uso de técnicas de secuenciación de nueva generación: miRNASeq. Estos datos proceden del trabajo publicado por Camps y colaboradores (Camps, Saini et al. 2014) y obtenidos del repositorio GEO (del inglés *Gene Expression Omnibus)* con identificador GSE47602. Concretamente, este experimento evalúa la expresión de miARNs en la línea celular de carcinoma de mama MCF7 bajo distintas condiciones de hipoxia: 0 (muestras usadas como control), 16, 32 y 48 horas en presencia de un 1% de oxigeno. En este caso, miARma-Seq identificó un total de 554 miARNs expresados, que representan un 91.8% de similitud con respecto al trabajo original. Por tanto, el resultado obtenido muestra gran relevancia, ya que en el caso de Camps, la anotación empleada procedía de una versión de miRBase (Kozomara and Griffiths-Jones 2014) diferente a la empleada en nuestro análisis. Además, en nuestro estudio, tal y como aconsejan diversos autores como el Dr. Anders (Anders, McCarthy et al. 2013) creímos conveniente eliminar todos los miARNs que presentaban bajos niveles de expresión. A continuación, se llevó a cabo el análisis de expresión diferencial (DE) de las tres condiciones experimentales frente a las muestras control, que proporcionó nuevamente, resultados con una alta correlación a los publicado previamente.

La correlación de Pearson de los valores logFC entre los datos de Camps y colaboradores y los obtenidos por miARma-Seq a las 16 horas fue de 0.97 (P.val = 0.0076); 0.99 a las 32 horas (P.val < 2.2e-16); y 0.98 a los 48 horas en condiciones de hipoxia (P.val < 2.2e-16). Asimismo, miARma-Seq identificó correctamente el 100% de los miARNs DE validados en el laboratorio por qPCR (PCR cuantitativa) alcanzando un valor de cambio de expresión muy similar al conseguido experimentalmente (Figura 17).

Por lo tanto, se confirmó que miARma-Seq es una herramienta válida para analizar experimentos de miRNASeq.

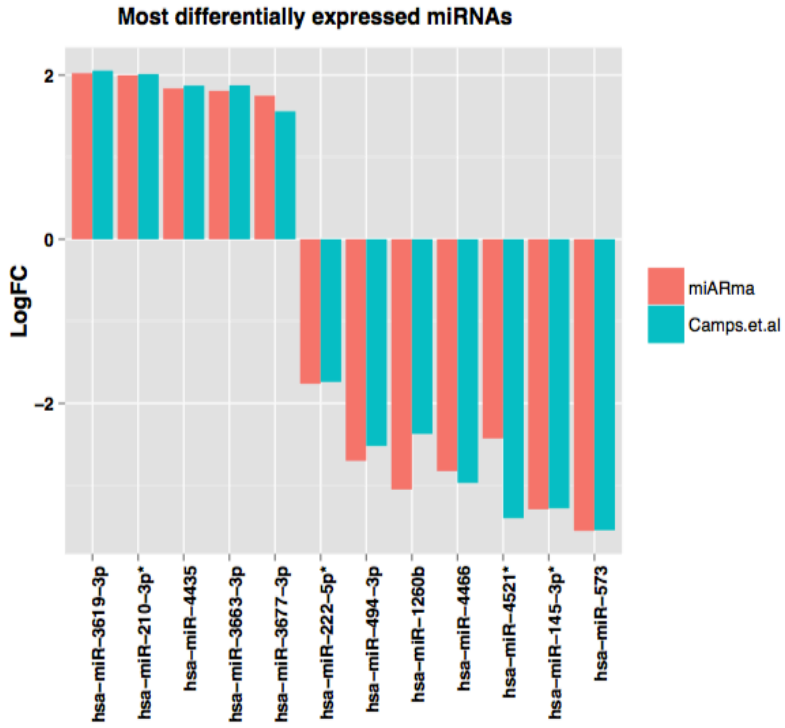


Figura 17. Estudio comparativo de logFC entre experimentos de miARNs.

Se representan los 5 miARNs más reprimidos y los 5 más sobre-expresados entre los resultados de Camps y colaboradores y miARma-Seq. Además, se indican los 3 miARNs validados en la publicación, que incluyen un *.

LogFC: log en base 2 del cambio de expresión entre las muestras control y las muestras bajo condiciones de hipoxia (oxígeno 1%).

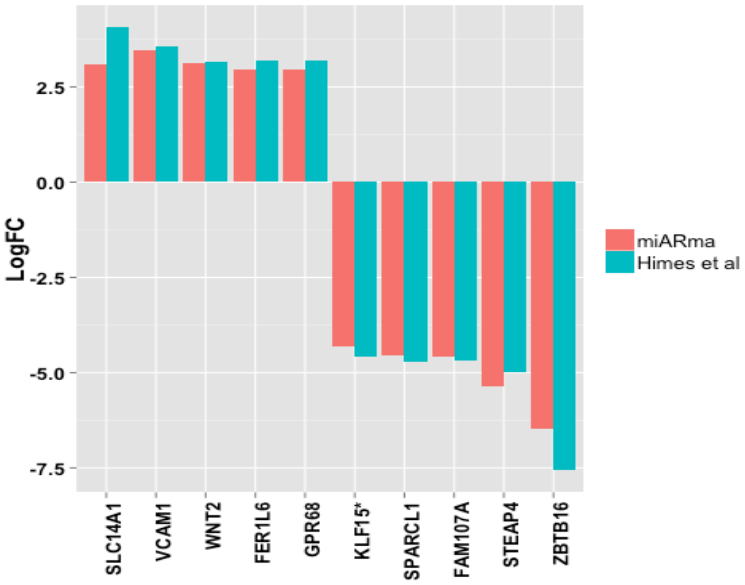
4.2.2. Validación de datos de expresión de genes.

El módulo de análisis de los niveles de expresión de genes implementado en miARma-Seq se evaluó, de igual forma, mediante la comparación de los resultados obtenidos por miARma-Seq frente a los de un trabajo con validación experimental incluida. En este caso, se estudió la expresión diferencial de genes procedente de diversas líneas celulares primarias de músculo liso, derivadas de vías respiratorias tratadas con dexametasona (Himes, Jiang et al. 2014), un potente glucocorticoide sintético que reduce la inflamación en pacientes diagnosticados de asma. Estos datos se obtuvieron de GEO (GSE37376).

Concretamente, miARma-Seq identificó un total de 1052 genes diferencialmente expresados (FDR < 0.05) entre las muestras control y las tratadas con dexametasona. El porcentaje de genes identificados por miARma-Seq en comparación con el trabajo original, alcanzó un valor de 98.48%, y la correlación entre los valores de expresión de ambos estudios un 0.99 (P.val < 2.2e-16). En la Figura 18 se muestra la elevada similitud obtenida de los valores de cambio de expresión, donde incluso entre los cinco genes más desregulados del experimento, encontramos resultados casi idénticos entre ambos análisis.

Figura 18. Análisis comparativo de logFC entre experimentos de ARNsm.

Se muestran los 5 ARNms más reprimidos y los 5 genes más sobre-expresados entre los resultados de Himes y colaboradores y miARma-Seq.
LogFC: log en base 2 del cambio de expresión entre las muestras control y las muestras bajo condiciones de hipoxia (oxígeno 1%).



Después de obtener una importante similitud en los resultados, incluso utilizando softwares distintos, evaluamos los datos proporcionados por nuestra herramienta, frente al conjunto de datos validados experimentalmente en el trabajo original. Como muestra la Figura 19, los valores de cambio de expresión logFC obtenidos por miARma-Seq y los comprobados experimentalmente mediante PCR cuantitativa presentaban una elevada semejanza.

En resumen, tras la evaluación de los datos obtenidos mediante el análisis de expresión diferencial de muestras de transcriptómica de microARNs, y de la información aportada por miARma-Seq tras el procesamiento de los datos brutos de un experimento de expresión de genes, se ha demostrado la elevada fiabilidad que presenta nuestra herramienta miARma-Seq en el análisis de muestras obtenidas mediantes técnicas de secuenciación de nueva generación. Además, la gran similitud obtenida entre nuestros resultados y los obtenidos por técnicas de validación experimental, confirmó la validez de miARma-Seq para el procesamiento de este tipo de muestras.

4.3. Análisis del conjunto de muestras de expresión del TCGA.

Tras la evaluación a la que fueron sometidas las herramientas necesarias para el procesamiento de datos procedentes de técnicas de NGS y el establecimiento de interacciones fiables, como son: miRGate y miARma-Seq, nos dispusimos a analizar un conjunto de 18.605 muestras obtenidas de un total de 15 tipos de tumores distintos tal y como se mostró en la Tabla 2.

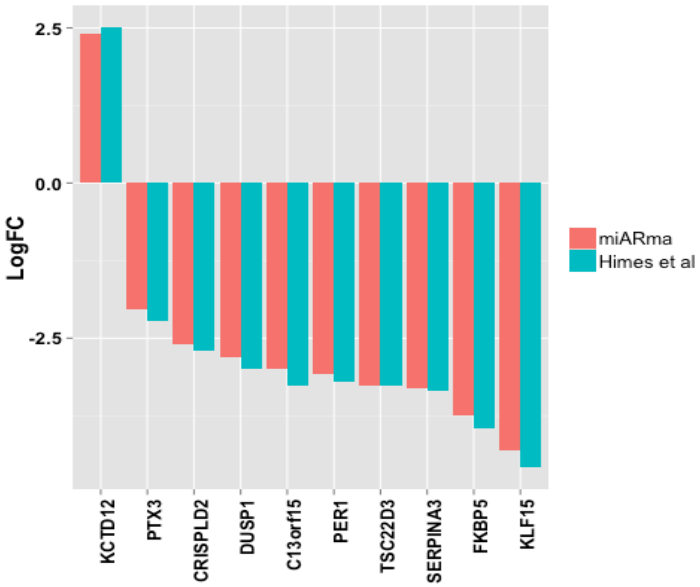


Figura 19. Comparación de logFC de aquellos genes validados experimentalmente.

Se indica el valor de expresión obtenido por q-PCR en el trabajo de Himes y colaboradores comparado con el dato obtenido in silico por miARma-Seq. Entre los genes validados aparece KCTD12 como único gen sobre-expresado, mientras 9 genes aparecen reprimidos.
LogFC: log en base 2 del cambio de expresión entre las muestras control y las muestras bajo condiciones de hipoxia. (oxígeno 1%).

4.3.1 Análisis de tumores individuales.

Cada uno de los 15 tumores se analizó siguiendo el protocolo recomendado (Anders, McCarthy et al. 2013) tanto para la obtención de genes y de miARNs diferencialmente expresados. Solo aquellos genes y/o miARNs con un FDR <0.05 y un cambio de expresión absoluto entre muestras control y muestras tumorales superiores 1, fueron seleccionados para estudios posteriores de regulación. La Tabla 6 muestra el total de genes y miARNs diferencialmente expresados resultantes, para cada uno de los tumores y según el tipo de muestra.

En esta tabla podemos observar como el tumor renal de célula clara y el tumor de vejiga, contienen el mayor número de genes y miARNs diferencialmente expresados, 334 y 339 respectivamente. Por el contrario, el tumor de tiroides y de próstata según los resultados obtenidos por miARma-Seq, obtienen el menor número de genes y miARNs desregulados, 61 genes ambos y 94 y 97 miARNs correspondientemente.

Asimismo, también se realizó un estudio de enriquecimiento funcional con el objetivo de identificar si la mayoría de los genes diferencialmente expresados pertenecían a alguna de las rutas relacionadas con el cáncer o si por el contrario, alguna de las rutas seleccionadas mostraba un número estadísticamente bajo de genes desregulados. Para ello en la Figura 20 se muestran la razón de probabilidades (para más detalles, consultar métodos) para aquellas rutas significativas (Pvalue ajustado<= 0.05). Cabe destacar, que con la única excepción del tumor de esófago (ESCA), todos los tipos de tumores incluidos presentan un empobrecimiento en genes diferencialmente expresados vinculados a la ruta de apoptosis.

Este resultado va en consonancia con lo expuesto en la introducción, donde se comentó que uno de las características del cáncer era la evasión a la muerte celular y de esta forma no sorprende obtener que la expresión de los genes implicados en esta ruta sean similares a los de una célula normal.

		RNASeq		miRNASeq	
Tipo de tumor	**Acrónimo**	**Genes +**	**Genes -**	**miARNs +**	**miARNs -**
Cromófobo de riñón	KICH	82	44	89	62
Carcinoma escamoso de cuello y cabeza	HNSC	138	16	165	103
Carcinoma Esofágico	ESCA	190	18	110	54
Carcinoma de riñón de célula papilar	KIRP	122	21	97	107
Carcinoma hepático	LIHC	170	22	150	31
Carcinoma de riñón de célula clara	KIRC	202	132	109	65
Adenocarcinoma de pulmón	LUAD	173	25	228	74
Carcinoma de Tiroides	THAD	38	8	61	33
Adenocarcinoma de próstata	PRAD	63	16	61	36
Carcinoma urotelial de vejiga	BLCA	182	46	326	73
Carcinoma invasivo de mama	BRCA	183	29	191	73
Carcinoma escamoso de pulmón	LUSC	153	49	237	84
Adenocarcinoma de estómago	STAD	144	31	121	69
Cholangiocarcinoma	CHOL	215	45	124	73
Carcinoma endometrial de Útero	UCEC	221	71	301	134

Tabla 6. Número de genes y miARNs diferencialmente expresados por tumor. La tabla muestra tanto los genes diferencialmente sobre-expresados (Genes +) y reprimidos (Genes -) del conjunto total de genes que participan en las siete rutas relacionadas con el cáncer. A su vez muestra el total de miARNs desregulados. Tanto los genes y miARNs han sido seleccionados por tener un FDR<0.05 y un cambio de expresión absoluto de logFC mayor a 1.

Los valores de razón de probabilidades y de probabilidad de cada ruta por cada tipo de tumor, pueden consultarse en la Tabla Suplementaria 2 del Anexo I.

A continuación se detallan brevemente los resultados obtenidos en cada tipo de tumor, dado que estos resultados son claves para su posterior integración en un estudio pan-cáncer.

4.3.1.1 Carcinoma urotelial de vejiga (BLCA).

El cáncer de vejiga urotelial es el tipo más común de cáncer de vejiga, se desarrolla en el revestimiento interno de la pared de la vejiga y se caracteriza por ser más común entre los hombres. Si se diagnostica tras la metástasis, sólo el 5% de los pacientes vivirá cinco años. De este tipo de tumor, se analizaron 433 muestras de RNASeq y, tal y como muestra la

Tabla 6, un total de 228 genes diferencialmente expresados fueron obtenidos. De los 182 genes sobre-expresados, el gen con expresión más exacerbada resultó ser *TUBA3C* perteneciente a la ruta de ciclo celular. Por el contrario, *PPP1R12B* (también denominado *MYPT2*), también del ciclo celular, resultó ser el gen más reprimido del conjunto de resultados. Aquellos genes y microARNs con valores mas extremos de cambio de expresión, se muestran en la Figura Suplementaria 1 del Anexo I.

Asimismo, se analizaron un total de 531 muestras de expresión de miRNASeq y 399 miARNs resultaron desregulados de forma significativa. miR-1 resultó ser el microARN más inhibido y miR-520f-3p el más expresado. El total de genes y microARNs diferencialmente expresados, sus valores del log2FC y de FDR, además de otros valores de interés, se muestran en las Tablas Suplementarias 3 y 4 del Anexo I.

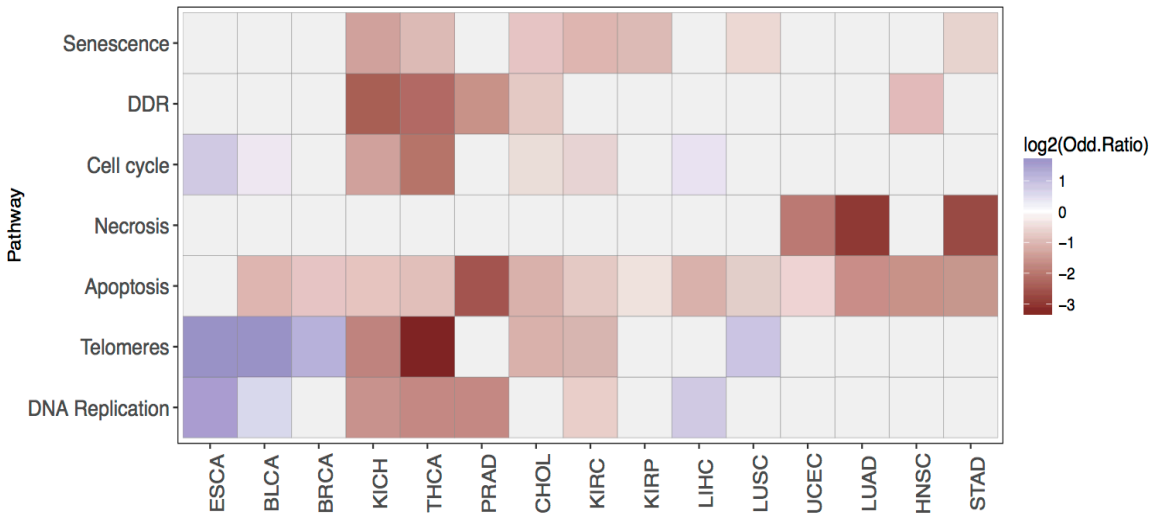


Figura 20. Enriquecimiento de los genes diferencialmente expresados en cada tipo de tumor, según su función. Razón de probabilidades (del inglés Odd Ratio, OR) de las rutas relacionadas con el cáncer significativas (P <0.05) en cada tipo de tumor. Los valores de OR se muestran en log2. En azul son las vías positivamente enriquecidas, mientras que en rojo, las negativamente enriquecidas (más alterado de lo esperado frente a menos alterado).

4.3.1.2 Carcinoma invasivo de mama (BRCA).

Este cáncer junto al de pulmón, es el más diagnosticado y la primera causa de muertes por cáncer en las mujeres españolas. De este tipo de cáncer, se analizaron 1247 muestras de expresión de genes y 2207 de microARNs, tal y como se detalla en la Tabla 2, 183 genes y 191 microARNs presentaban unos niveles de expresión estadísticamente más altos en muestras tumorales frente a muestras de pacientes sanos (Tabla 6). De entre ellos destacamos al gen *HIST1H2AJ*, una histona implicada en ciclo celular, elongación de

telómeros y senescencia, y al microARN miR-1269a. De forma contraria, obtuvimos 29 genes y 73 microARNs inhibidos en las muestras tumorales, tales como el gen *BMX*, implicado en apoptosis y el miARNs hsa-miR-206, ambos fuertemente reprimidos. Aquellos genes y microARNs con los cambio de expresión mas acusados, se muestran en la Figura Suplementaria 2 del Anexo I. Además, sus valores de log2FC, FDR, además de otros valores relevantes, se muestran en las Tablas Suplementarias 3 y 4 del Anexo I.

4.3.1.3 Cholangiocarcinoma (CHOL).

Cholangiocarcinoma es el termino que describe a los tumores que se originan en los conductos biliares que transportan la bilis desde el hígado y la vesícula biliar hasta el intestino delgado.. Se le considera como un cáncer raro, aunque la mortalidad asociada ha aumentado en todo el mundo en las últimas tres décadas. Al ser un tumor considerado raro, sólo se obtuvieron 45 muestras. De ellas se realizaron estudios de expresión de genes y de microARNs obteniendo 215 genes sobre-expresados y 45 inhibidos (Tabla 6). En función del cambio obtenido en los niveles de expresión, sobresalen por su alta expresión *HMGA2*, relacionado con senescencia y *GADD45G*, de ciclo celular por su fuerte represión. En el caso de los microARNs, se obtuvieron 124 microARNs altamente expresados, enfatizando a miR-526b-5p y, 73 estadísticamente reprimidos. De entre ellos subrayamos la elevada represión del miARN hsa-miR-5589-3p y de su pareja hsa-miR-5589-5p. Para más detalles sobre los genes y microARNs diferencialmente expresados, se pueden consultar las Tablas Suplementarias 3 y 4 del Anexo I, a su vez los genes y microARNs con grandes cambios de expresión, se muestran en la Figura Suplementaria 3 del Anexo I.

4.3.1.4 Carcinoma Esofágico (ESCA).

Este cáncer representa el 1% de todos los nuevos casos de cáncer en nuestro país. El tipo escamoso está relacionado con el tabaquismo y el consumo intensivo de alcohol y el adenocarcinoma se asocia a trastornos relacionados con el reflujo del ácido del estómago en el esófago y el esófago de Barret. Cerca de 400 muestras fueron analizadas, 197 de expresión de genes y 199 de microARNs (Tabla 2). De los datos de RNASeq obtuvimos 208 genes diferencialmente expresados, 190 con unos valores elevados de forma significativa, como *CCNA1*, un gen altamente sobre-expresado que forma parte del ciclo celular, senescencia y replicación del DNA. Por el contrario, *PPP1R12B/MYPT2*, del ciclo celular, resultó ser el gen más reprimido del total de los 18 genes obtenidos. En conjunto, los genes pertenecientes a las rutas de replicación del ADN, elongación de telómeros y

ciclo celular, presentan unos cambios de expresión muy significativos entre las muestras control y las de los pacientes.

Asimismo, obtuvimos 110 microARNs sobre-expresados y 54 inhibidos (Tabla 6). Entre estos miARNs estadísticamente significativos, podemos destacar a miR-372-3p por su elevada expresión y a miR-490-3p por sus bajos niveles de expresión. Estos microARNs, junto con los genes con mayor variación de expresión, se muestran en la Figura suplementaria 4 del Anexo I. El resto de detalles pueden consultarse en las Tablas Suplementarias 3 y 4 del mismo Anexo.

4.3.1.5 Carcinoma escamoso de cabeza y cuello (HNSC).

La mayoría de los cánceres de cabeza y cuello comienzan en las células escamosas que forman las membranas mucosas que recubren el interior de la boca, la nariz y la garganta.

Un total de 138 genes sobre-expresados y 16 genes reprimidos (Tabla 6) se obtuvieron tras el análisis de 511 muestras. El gen que resultó mas fuertemente expresado fue *SMC1B* perteneciente al ciclo celular y por el contrario, *MAPT* de la ruta de apoptosis, el más inhibido. Además, se analizaron 932 muestras de expresión de microARNs, obteniendo 165 miARNs sobre-expresados y 103 con una expresión estadísticamente reducida. El microARN con una mayor aumento de expresión obtenido fue miR-767-5p y miR-375 el más inhibido. Para más detalles sobre los genes y microARNs diferencialmente expresados, se pueden consultar las Tablas Suplementarias 3 y 4 del Anexo I, a su vez los genes y microARNs con grandes cambios de expresión, se muestran en la Figura Suplementaria 5 del Anexo I.

4.3.1.6 Tumor cromófobo de riñón (KICH).

El carcinoma de células renales cromófobas es un tipo raro de cáncer de riñón, por consiguiente, solo el 5% de los tumores de riñón diagnosticados en Estados Unidos, fue descrito como tumor cromófobo de riñón. De esta forma, solo se obtuvieron un total de 91 muestras entre controles (26) y pacientes enfermos (66), tanto de muestras de RNASeq como de miRNASeq. Tras el estudio de expresión diferencial, 88 genes resultaron tener unos valores de expresión sobre-expresados estadísticamente y 44 inhibidos (Tabla 6), de ellos podemos destacar a *CDKN2A* relacionado con ciclo celular y senescencia como gen con expresión mas exacerbada y a *PRKCQ*, de apoptosis, como gen más reprimido. Referente a los microARNs, 151 resultaron diferencialmente expresados, de los cuales 89 mostraron unos valores de expresión estadísticamente altos. De este grupo destaca miR-891a-5p y, por el contrario 62 mostraron una baja expresión (Tabla 6). De estos,

destacamos a miR-184 por ser el miARN mas reprimido de esta enfermedad. El total de genes y microARNs diferencialmente expresados además de otros valores de interés, se exponen en las Tablas Suplementarias 3 y 4 del Anexo I, a su vez la Figura Suplementaria 6 muestras los genes y microARNs con mayor cambio de expresión.

4.3.1.7 Carcinoma de riñón de célula clara (KIRC).

El tipo más común de cáncer de riñón se le denomina carcinoma de células renales y existen dos tipos principales, el de célula papilar y el de célula clara. El carcinoma de célula clara es el tipo más común, dado que representa aproximadamente el 80 por ciento de los carcinomas de riñón.

Más de mil muestras de expresión tanto de genes como de microARNs se analizaron de este tipo de tumor y se obtuvieron un total de 202 genes sobre-expresados y 132 reprimidos (Tabla 6). Asimismo, inferimos 109 microARNs altamente expresados y 65 inhibidos. Entre los genes, destacamos a por su elevado nivel de expresión en muestras tumorales a *TEXT15*, de ciclo celular y a *HSPA2*, de la misma ruta, por su baja expresión en este tipo de muestras. De forma similar, entre los microARNs subrayamos a miR-122-5p por su expresión exacerbada y a miR-508-5p y 3p, por su alta represión. Mas detalles en la Tabla Suplementaria 4 del Anexo I, así como en la Figura Suplementaria 7.

4.3.1.8 Carcinoma de riñón de célula papilar (KIRP).

Como acaba de comentarse, el tumor de célula papilar representa aproximadamente el 15 por ciento del total de tumores de riñón diagnosticados.

De este tipo de tumor fueron analizadas cerca de 700 muestras de expresión. En unas 300 muestras de RNASeq, pudimos obtener un conjunto de 122 genes sobre-expresados y 21 inhibidos (detalles en Tabla 6), entre ellos recalcamos a *BIRC7* de apoptosis y a *HSPA2*, al igual que en tumor de riñón de célula clara, de ciclo celular. Asimismo, aproximadamente 450 muestras de transcriptoma de microARNs fueron también analizadas y 97 y 107 microARNs resultaron sobre-expresados e inhibidos (Tabla 6), respectivamente. miR-9-5p resultó tener unos valores de expresión altamente elevados en muestras de pacientes frente a muestras sanas y con valores de expresión contrarios obtuvimos a miR-184. Para más detalles sobre los genes y microARNs diferencialmente expresados, se pueden consultar las Tablas Suplementarias 3 y 4 del Anexo I, a su vez los genes y microARNs con grandes cambios de expresión, se muestran en la Figura Suplementaria 8 del Anexo I.

4.3.1.9 Carcinoma hepático (LIHC).

El carcinoma hepatocelular es la forma más común de cáncer de hígado, constituyendo más del 80 % de los casos. La mayoría (78%) son debidos a infecciones crónicas de hepatitis B o C o cirrosis y representan la tercera causa de muertes relacionadas con el cáncer.

De la comparación de los niveles de expresión de los genes de las 50 muestras control, frente a las 268 tumorales, obtuvimos 170 genes altamente expresados en muestras de pacientes y 22 con unos niveles muy bajos (Tabla 6). El gen *TERT*, vinculado al ciclo celular y a la elongación de telómeros destacó por estar estadísticamente muy sobre-expresado. Por el contrario, *FOS* perteneciente a la ruta de senescencia, resultó ser el gen más reprimido.

En el caso de los microARNs, se analizaron mas de 500 muestras y 150 microARNs mostraron uno valores exacerbados estadísticamente y sin embargo, 31 miARNs presentaban unos valores de expresión muy reducidos (Tabla 6). Así, miR-767 se expresaba de forma muy elevada en muestras tumorales, mientras que miR-490-3p, se expresaba más en muestras control. Aquellos genes y microARNs desregulados además de otros valores del interés, se muestran en las Tablas Suplementarias 3 y 4 del Anexo I. Asimismo, la Figura Suplementaria 9 muestra los 5 genes y los 5 miARNs con mayores cambios en sus niveles de expresión.

4.3.1.10 Adenocarcinoma de pulmón (LUAD).

El cáncer de pulmón, de forma global, es responsable del mayor número de muertes tanto en hombres y mujeres, dado que representa el 23 por ciento de todas las muertes debidas al cáncer. Dentro de los diferentes tipos de cáncer de pulmón, el tipo más común es llamado cáncer de pulmón de células no microcíticos (microcítico: células pequeñas). Específicamente, los subtipos que se estudian se denominan adenocarcinoma de pulmón y carcinoma de células escamosas de pulmón. El adenocarcinoma representa el 30% de los carcinomas no microcíticos y es el menos relacionado con el consumo de tabaco.

Un total aproximado de 600 muestras de ARN de pacientes con adenocarcinoma de pulmón y controles, fueron analizadas y 173 genes sobre-expresados y 25 reprimidos resultaron del estudio (Tabla 6). La histona *HIST1H1E*, relacionada con senescencia y apoptosis, resultó ser el gen más sobre-expresado, *IL6* de senescencia, por el contrario, el gen más reprimido. Asimismo, más de 1000 muestras de ARNs de pequeño tamaño fueron procesadas, y 228 y 74 microARNs (Tabla 6), resultaron mostrar unos valores de expresión altos y disminuidos, respectivamente. Entre ellos merece la pena destacar al miR-372-3p

por su alta sobre-expresión y al miR-3168 por su fuerte inhibición. Mas detalles se pueden consultar las Tablas Suplementarias 3 y 4 del Anexo I, a su vez los genes y microARNs con grandes cambios de expresión, se muestran en la Figura Suplementaria 10 del Anexo I.

4.3.1.11 Carcinoma escamoso de pulmón (LUSC).

La variedad del carcinoma escamosos de pulmón, es el cáncer broncopulmonar más frecuente en nuestro país, representando el 40% de los carcinomas no microcíticos.

Quinientas cuarenta muestras de RNASeq se analizaron y obtuvimos 202 genes diferencialmente expresados (Tabla 6), 153 sobre-expresados, como *DSG3* de apoptosis, el más fuertemente expresado y 49 inhibidos, como *TUBB1* de ciclo celular, como el más reprimido. De las 915 muestras de miRNASeq, obtuvimos 237 microARNs sobre-expresados, como miR-372-3p, el miARNs más sobre-expresado tanto en adenocarcinoma como en carcinoma escamoso de pulmón y 84 con baja expresión en muestras tumorales, como miR-4529-5p, el más reprimido. Estos microARNs, junto con los genes con mayor variación de expresión, se muestran en la Figura suplementaria 11 del Anexo I. El resto de detalles pueden consultarse en las Tablas Suplementarias 3 y 4 de mismo Anexo.

4.3.1.12 Adenocarcinoma de próstata (PRAD).

Este tipo de cáncer se desarrolla en la próstata, una glándula que se localiza en el sistema reproductivo masculino y representa el tumor más diagnosticado en hombres. Casi todo el cáncer de próstata diagnosticado es del tipo adenocarcinoma. El grado de desarrollo de esta enfermedad se clasifica en función de la escala de Gleason, una puntuación de Gleason baja, significa que el tejido del cáncer es similar a las células, por el contrario, una valor elevado significa que las células cancerosas son muy diferentes de las células normales y es probable que se propaguen.

Aproximadamente 400 muestras de expresión de genes entre controles y pacientes diagnosticados de adenocarcinoma de próstata, fueron analizados. 79 genes diferencialmente expresados se obtuvieron (Tabla 6), donde destacó *UNC5A*, vinculado a apoptosis, entre los 63 genes más sobre-expresados y *UNC5B*, también de apoptosis, entre los 16 genes inhibidos que se obtuvieron. En el caso de los microARNs, se procesaron 737 muestras obteniendo 61 miARNs sobre-expresados y 36 reprimidos. Como microARNs altamente expresado, obtuvimos a miR-449a y como más inhibido a miR-891a-5p. Para mas detalles se puede consultar las Tablas Suplementarias 3 y 4, así como la Figura Adicional 12 del Anexo I.

4.3.1.13 Adenocarcinoma de estómago (STAD).

El cáncer de estómago, es el 6º cáncer más común en España. La incidencia varía ampliamente dependiendo de factores genéticos y ambientales, como el tabaquismo, la obesidad y una dieta alta en alimentos ahumados, junto con infecciones ya sea por Helicobacter pylori o por el virus Epstein-Barr.

Se realizó un análisis de expresión diferencial entre las 450 muestras de ARN obtenidas y se obtuvieron 175 genes estadísticamente significativos, 144 fuertemente expresados y 31 inhibidos (Tabla 6). *BIRC7* asociado a apoptosis, resultó ser el gen más sobre-expresado, por el contrario *PKP1* de la misma ruta, fue el más reprimido.

En el caso de los miARNs, el más expresado entre las cerca de 800 muestras analizadas, aparece miR-372a-3p, al igual que en ambos tumores de pulmón y como microARN más fuertemente reprimido a miR-6510-3p. Aquellos genes y microARNs desregulados, se muestran en las Tablas Suplementarias 3 y 4 del Anexo I. Asimismo, la Figura Suplementaria 13 muestra los 5 genes y los 5 miARNs con mayores cambios de expresión.

4.3.1.14 Carcinoma de Tiroides (THCA).

El cáncer de tiroides se desarrolla en una glándula en la parte frontal del cuello por debajo de las cuerda vocales. El carcinoma papilar de tiroides, el tipo que aquí se estudia, es el tipo más común de cáncer de tiroides, ya que actualmente representa el 80 % de los casos diagnosticados.

Un total de 59 muestras control y 498 muestras tumorales de transcriptómica génica fueron estudiadas en detalle y apenas 38 genes aparecieron sobre-expresados y 8 reprimidos (Tabla 6). *BIRC7* de apoptosis, como ya obtuvimos en el tumor de estómago, resultó ser el gen más sobre-expresado y por el contrario, *JUN*, asociado a senescencia, como el más inhibido. En el estudio de los microARNs, de más de 1000 muestras analizadas, 61 miARNs presentaron un incremento de su nivel de expresión en las muestras tumorales, sin embargo 33 microARNs mostraron una disminución al compararse con las muestras control, tales como el sobre-expresado miR-146b-3p y miR-1258, el microARNs más inhibido. Para más detalles, se puede consultar la Tabla Suplementaria 4 y la Figura Adicional 14 del Anexo I.

4.3.1.15 Carcinoma endometrial de Útero (UCEC).

El cáncer de endometrio se desarrolla en las células que forman el revestimiento interno del útero o el endometrio, y es uno de los cánceres más comunes del sistema reproductivo femenino entre las mujeres.

De este tipo de carcinoma, se analizaron 208 muestras de expresión de genes y 1097 de microARNs, tal y como muestra la Tabla 2. 221 genes y 301 microARNs presentaban unos niveles de expresión estadísticamente más altos en muestras tumorales frente a muestras de pacientes sanos (Tabla 6). De entre ellos destacamos al gen *SFN*, implicado en ciclo celular, y apoptosis, y al microARN miR-1269a-3p. De forma opuesta, obtuvimos 71 genes y 134 microARNs inhibidos en las muestras tumorales, tales como el gen *PPP1R12B/MYPT2*, implicado en ciclo celular y el miARN hsa-miR-202-5p, ambos fuertemente reprimidos. Más detalles sobre el total de genes y miARNs desregulados de este tumor, pueden examinarse en las Tablas Suplementarias 3 y 4 así como la Figura Suplementaria 15 del Anexo I.

4.3.2 Análisis integrado del conjunto de tumores.

Una vez fueron analizadas las 18.605 muestras procedentes de los 15 tipos de tumores (Tabla 2), se seleccionaron aquellos genes y microARNs diferencialmente expresados (FDR <=0.05) con un cambio significativo de expresión entre las muestras control y las muestras tumorales (valor absoluto de log2FC >= 1) resultantes de cada tipo de tumor, tal y como acaba de mencionarse en el apartado anterior (Tabla 6). Como primer paso para realizar un estudio de tipo pan-cáncer, estudiamos todos los genes que mostraban unos patrones de expresión diferencial similar, en al menos ocho tipos tumorales diferentes, dado el conjunto de genes obtenido a este nivel de corte, presentaba un alto enriquecimiento en genes con conocida implicación en el cáncer según COSMIC (OR 3.56 p<1.54E-2) y NCG (OR 2.8, p<1.87E-3). Para más detalles, consultar la Tabla 3a y la sección de métodos.

Del total de 147 genes diferencialmente expresados en al menos 8 tipos tumorales (Tabla Suplementaria 5 del Anexo I), identificamos el gen *SPC24*, perteneciente al ciclo celular, como el único gen sobre-expresado en todos los tipos de tumores estudiados. También se identificó a *SYNE1* (igualmente del ciclo celular) por ser el gen con los niveles de expresión más reprimidos en los 15 tumores estudiados (véase la Figura 21 para la distribución de *SPC24* y *SYNE1*). Además, inferimos que 11 genes pertenecen a la ruta de apoptosis, donde *E2F1* y Claspin (*CLSPN*) muestran una perfil de desregulación conservado en un total 14 y 12 tipos de tumor, respectivamente. Asimismo, del resto de rutas, podemos destacar los 126 genes diferencialmente desregulados en el ciclo celular y 21 procedentes de la ruta de respuesta al daño en el ADN (DDR), donde los oncogenes *CDT1*, *PLK1* y *DTL* se encuentran entre los genes más destacados. También, de la vía de

replicación del ADN, obtuvimos 26 genes desregulados, incluyendo a los ya mencionados *E2F1*, *CDT1* y al oncogén *GINS2*. De los 29 genes alterados procedentes de la vía de senescencia, podemos destacar a varios genes que aparecen sobre-expresados en al menos 14 de los 15 tipos de tumores estudiados, como *Ubhc10/UBE2C*, *E2F1* o *CCNA2*. Finalmente, las polimerasas *DNA2* y *POLE2* que se encuentran sobre-expresadas en 12 tipos de tumores, destacan de la ruta de elongación de telómeros.

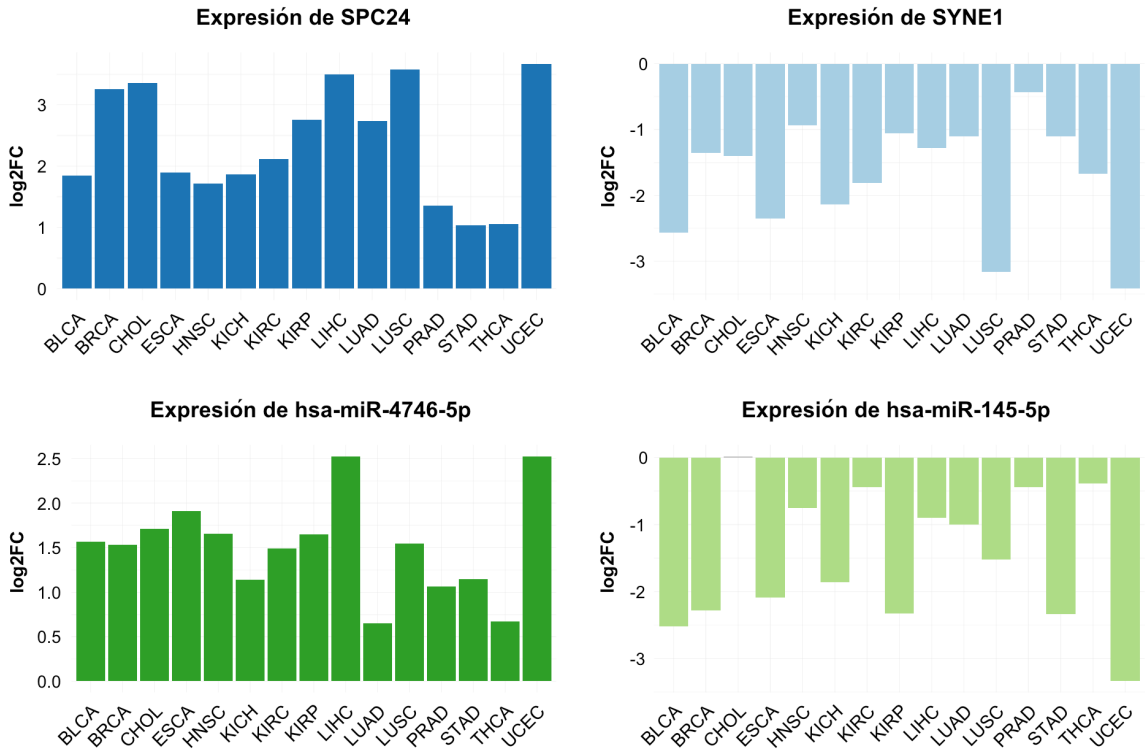


Figura 21. Genes y microARNs diferencialmente expresados en todos los tumores estudiados. SPC24, aparece sobre-expresado en los 15 tipos de tumores estudiados. SYNE1 por el contrario, se muestra reprimido en todos los tipos de cáncer estudiados; miR-4746-5p se expresa de manera exacerbada en los 15 tipos de tumores, mientras que miR-145-5p, se encuentra inhibido en todos los tumores excepto Cholangiocarcinoma (CHOL).

Igualmente, estudiamos los microARNs diferencialmente expresados que obtuvimos tras el análisis realizado al conjunto de los 15 tipos de tumores. En consecuencia, tal y como se detalla en la Tabla Suplementaria 6 del Anexo I, extrajimos 95 miARNs desregulados estadísticamente, en al menos 8 tipos diferentes de tumores dado que a este nivel de corte, obtuvimos un mayor enriquecimiento en microARNs relacionados con el cáncer según la base de datos OncomirDB (OR 8.13, p< 5.25E-18, Tabla 3b). Entre estos, destacamos el clúster oncogénico formado por miR-182, miR-96 y miR-183, que aparece sobre-expresado en 14 tipos de tumores. También miR-4746, que tal y como se muestra en la Figura 21, se

expresa de forma muy destacada en las muestras tumorales independientemente del tipo de tumor. Por el contrario, los anti-oncomiRs miR-145, miR-139 y miR-195 fueron los miARNs más fuertemente reprimidos en nuestro conjunto de datos. Curiosamente, miR-145 y miR-143 (que también aparece inhibido), pertenecen a clúster supresor de tumores miR-143/145 (ver Figura 21) implicado en los pasos iniciales de la tumorigenesis (Cordes, Sheehy et al. 2009). Por otra parte, también aparecen otros microARNs supresores de tumores fuertemente reprimidos, como miR-195 y miR-139 (Yonemori, Seki et al. 2016).

4.3.2.1. Interacciones miARN-ARNm conservadas entre diferentes tipos tumorales.

Los 147 genes y 95 microARNs diferencialmente expresados en la mayoría de los tipos de tumores incluidos en este trabajo (detallados en las Tablas 5 y 6 del Anexo I.) fueron recopilados y usados para estudiar las posibles redes de regulación establecidas entre ellos, mediante el uso de miRGate (Andres-Leon, Gonzalez Pena et al. 2015). Debido a que el mecanismo de acción mas aceptado de los microARNs es la regulación de la expresión génica, reduciendo los niveles de ARN mensajero, miRGate calculó las asociaciones, en función de aquellos genes y miARNs con valores de expresión inversamente correlacionados en los distintitos tipos de tumores. Por último, con la intención de aumentar aún más la fiabilidad de los resultados, sólo se consideraron aquellas predicciones que estuvieran respaldadas por más de un método de predicción siempre y cuando dicho sitio de unión, correspondiera a las mismas coordenadas genómicas del 3'-UTR (para detalles ver métodos). Como resultado, se obtuvo un conjunto de 41 interacciones (Tabla Suplementaria 7 del anexo I).

Posteriormente, analizamos la correlación de la expresión entre los genes y microARNs que forman cada una de las 41 asociaciones obtenidas. Para ello se creó un modelo de regresión lineal que incluía la expresión del gen (logaritmo de las RPKMs) como variable respuesta y, la expresión del miARN (logaritmo del las RPKMs), además de la metilación del gen (MET) y la alteración en el número de copias (CNAs), como variables predictoras para cada tipo tumoral (Tabla Suplementaria 9 del Anexo I). Dado el bajo número de muestras en algunos tipos tumorales como el tumor cromófobo de riñón o el cholangiocarcinoma, se realizó además, un análisis conjunto, formado por todas las muestras tumorales. Con el objetivo de tener en cuenta la diferente expresión de los genes y miARNs en los distintos tipos tumorales, el tipo de cáncer y la posible interacción con la expresión del microARN fueron asimismo incluidos, como factores adicionales (ver Métodos). Treinta seis de las 41 interacciones obtenidas (Figura 22, Tabla Suplementaria 7

del anexo I), mostraron una correlación negativa estadísticamente significativa tras el análisis de regresión conjunto (Figura 23).

De las asociaciones inferidas, 25 de ellas son nuevas, dado que no han sido descritas en ningún otro trabajo científico de acuerdo con diversas bases de datos especializadas, como son OncomirDB (Wang, Gu et al. 2014), miRCancer (Xie, Ding et al. 2013) y miRTarBase (Chou, Chang et al. 2016). De estas nuevas interacciones, una mayoría están relacionadas con el ciclo celular, como las obtenidas por los genes *SPC24* y *CDC20* regulados por el anti-oncomiR miR-139, y *PKMYT1* regulado por el supresor tumoral miR-195. Referente a la vía de senescencia, la mayoría de las asociaciones que hemos obtenido fueron descubiertas previamente por otros autores, aunque entre las nuevas interacciones podemos destacar las formadas por *HMGA2* / miR-139 y *EZH2* / miR-195. Asimismo, en la ruta encargada de la replicación del ADN, la interacción formada por el gen *MCM2* y los anti-oncomiRes miR-139 y miR-143, destaca al conservarse en la mayoría de los tipos tumorales. Finalmente, las asociación entre *PRKAR2B* / miR-135b destacó por aparecer en un alto número de tipos de cáncer y relacionado con la apoptosis.

Una problemática existente a la hora de diseñar terapias de restablecimiento de expresión en función de genes y microARNs desregulados, son los efectos secundarios de regulación inesperados. Esto es debido a que los miARNs pueden regular cientos de genes diferentes y por lo tanto a su vez, un gen puede estar regulado por diferentes microARNs. Para reducir esta limitación, hemos calculado un conjunto de interacciones de "alta especificidad", formada por los genes y microARNs diferencialmente expresados, que presentan, un número muy bajo de interacciones secundarias (más detalles en la sección de métodos), y que de esta manera, su baja promiscuidad pueda disminuir los efectos colaterales adversos.

De los datos procesados, obtuvimos 17 pares significativos ($p \leq 0.05$), formados en gran parte por miARNs supresores de tumores (por ejemplo, miR-let-7c, miR-139, miR-145 o miR-195) asociados a genes formadores de tumores, u oncogenes, como aquellos pertenecientes al ciclo celular o a la vía de respuesta al daño en el ADN: *BUB1B, AURKA, BIRC5, CENPK*, *BRCA1* y *CHEK1*. Once de los 17 pares obtenidos no han sido previamente identificados (Más detalles en la Figura 24 y en la Tabla 8 del Anexo I) y destacamos el interés de estas interacciones dado su posible potencial para el desarrollo de fármacos basados en oligonucleótidos anti-microARNs (AMOs).

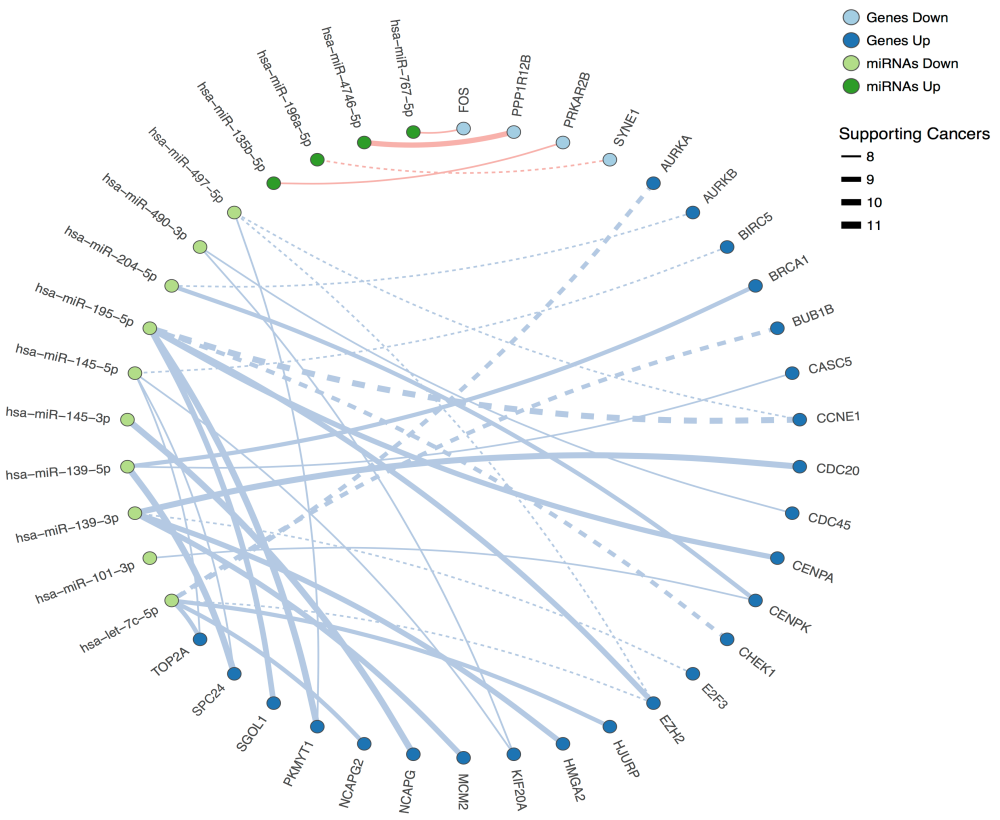


Figura 22. Interacciones miARN-ARNm relevantes en la mayoría de los tumores estudiados. Interacciones entre miARNs (verde) y genes (azul). La línea continua indica interacciones nuevas, mientras que las líneas discontinuas indican interacciones propuestas por otras autores y validadas experimentalmente. Las líneas rojas conectan miARNs Up con genes reprimidos.

Una vez estudiadas las interacciones existentes entre la mayoría de los tipos de cáncer, comprobamos que entre los dos tipos tumorales procedentes de pulmón (adenocarcinoma o LUAD y tumor escamoso de pulmón o LUSC) aparecían un conjunto de interacciones exclusivas y que por lo tanto no aparecían en el resto de tipos tumorales. Para estos 79 pares, creamos modelos de regresión lineal para tener en cuenta el efecto de la metilación y de las CNAs, de la misma forma que se empleó en los pares conservados (Tabla Suplementaria 11 del Anexo I). Como resultado encontramos 40 interacciones formadas por 35 genes diferencialmente expresados (sólo 4 reprimidos, véase la Figura 25) y 13 microARNs estadísticamente desregulados. De ellos destacó miR-1976, dado que de las 40 asociaciones, 16 están formadas por este microARN reprimido. De entre ellas, 16 genes están implicados en el ciclo celular, la apoptosis y la ruta DDR. El segundo miARN más representado en estos pares exclusivos fue el supresor de tumores let-7b-5p, actuando como

regulador en 10 interacciones, de las cuales 8 están formadas por genes implicados en el ciclo celular y 2 en senescencia.

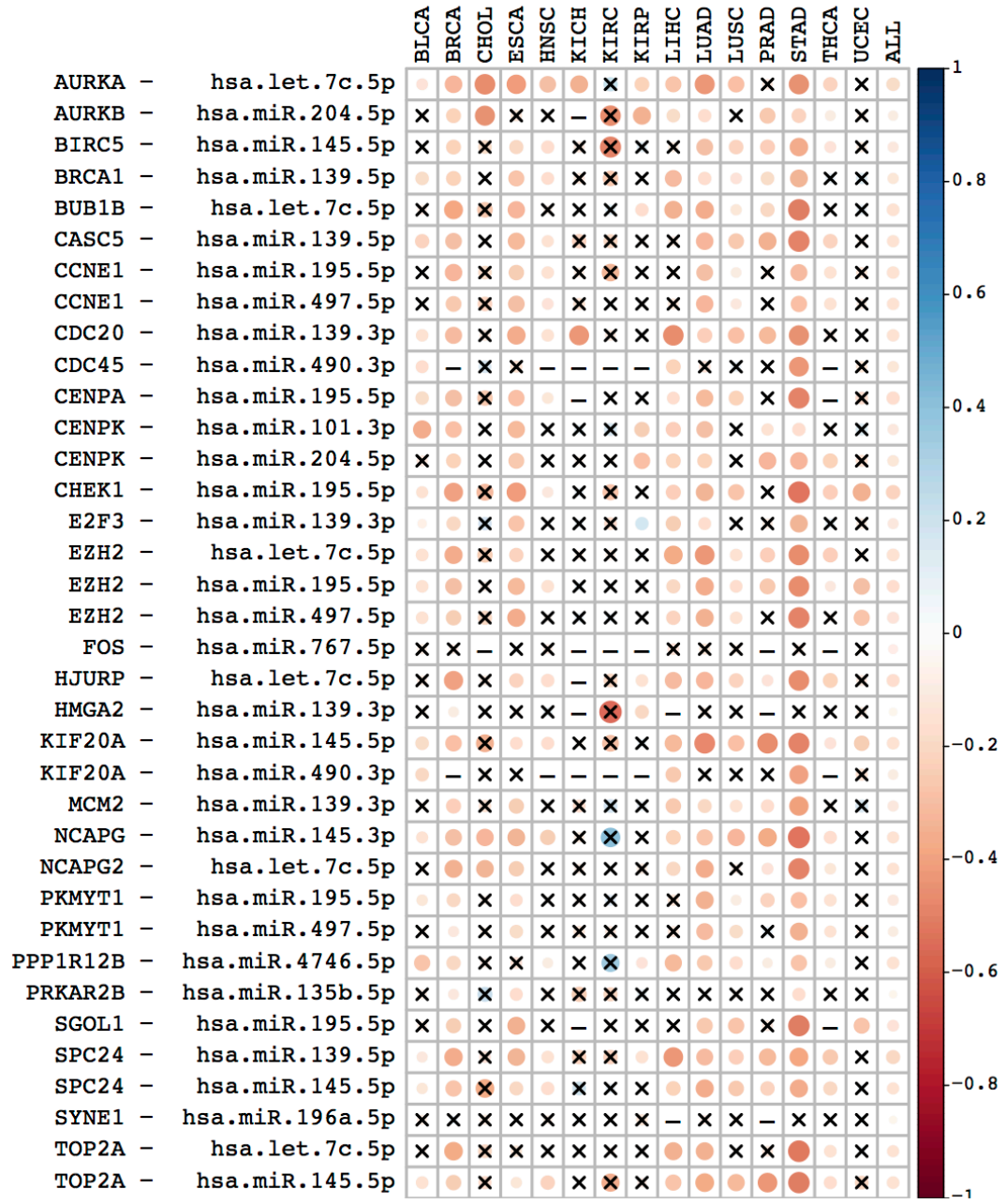


Figura 23. Correlación entre la expresión del gen y del microARN de cada interacción conservada. Correlación mediante el uso de un modelo de regresión linean entre la expresión del gen y la del microARN junto a la metilación y los CNAs para cada par (fila) en cada tipo tumoral (columna) y al conjunto de muestras ALL. El tamaño del circulo es proporcional al valor absoluto de la correlación, correlaciones negativas en rojo y positivas en azul. Aquellos modelos que no alcanzan una significancia estadística (FDR-corregido <0.05) se marcan con una equis. Los modelos donde alguno de los integrantes, gen o miARN no es detectado se marcan con un guion.

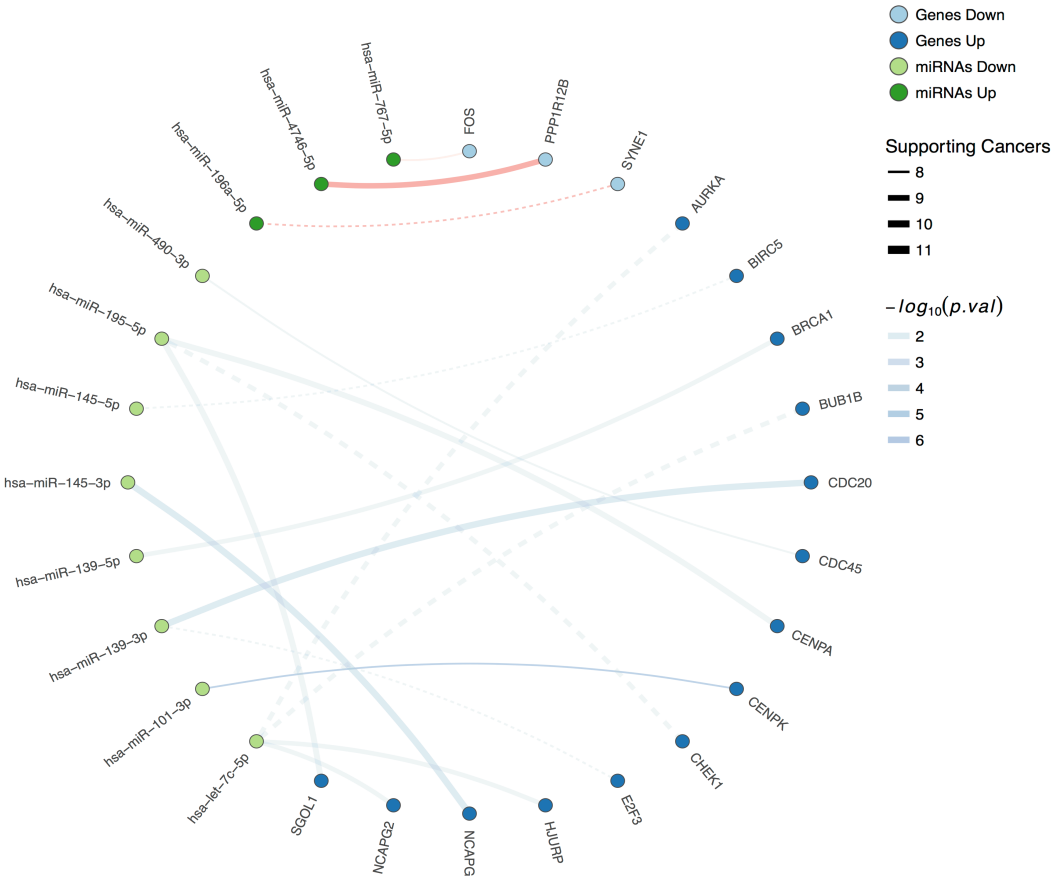


Figura 24. Interacciones entre miARNs y genes específicas. Interacciones entre miARNs (verde) y genes (azul). La línea continua indica interacciones nuevas, mientras que las líneas discontinuas indican interacciones validadas por otros autores. Se incluye el P-valor ajustado.

4.3.2.2. Análisis de supervivencia.

Para finalizar, se exploró la relación entre los niveles de expresión de los genes procedentes de las interacciones descubiertas, con la supervivencia de los pacientes. De esta forma, para los pares conservados en la mayoría de los tumores, se estudió el efecto de la expresión del gen en cada uno de los tipos tumorales, mediante el uso de los modelos de riesgo proporcionales de Cox. En estos modelos, tanto el valor de expresión del gen (RPKMs transformados logarítmicamente), como el estadío tumoral (codificado en dos categorías, estadíos I-II frente a estadíos III y IV) fueron usados como variables predictoras. En el caso del tumor de próstata (PRAD), debido a la ausencia de información referente a los estadíos, se empleó el grado tumoral como variable predictora (Valores de Gleason 6-7 frente a 8-10). Aquellos pares que exhiben una correlación significativa entre la expresión del gen y de su microARN regulador, junto a una asociación independiente del estadío tumoral, entre la expresión génica y la supervivencia del paciente, apoyan la hipótesis de que el

microARN regula la expresión del gen y esta alteración, es la que posteriormente afecta a la supervivencia del paciente.

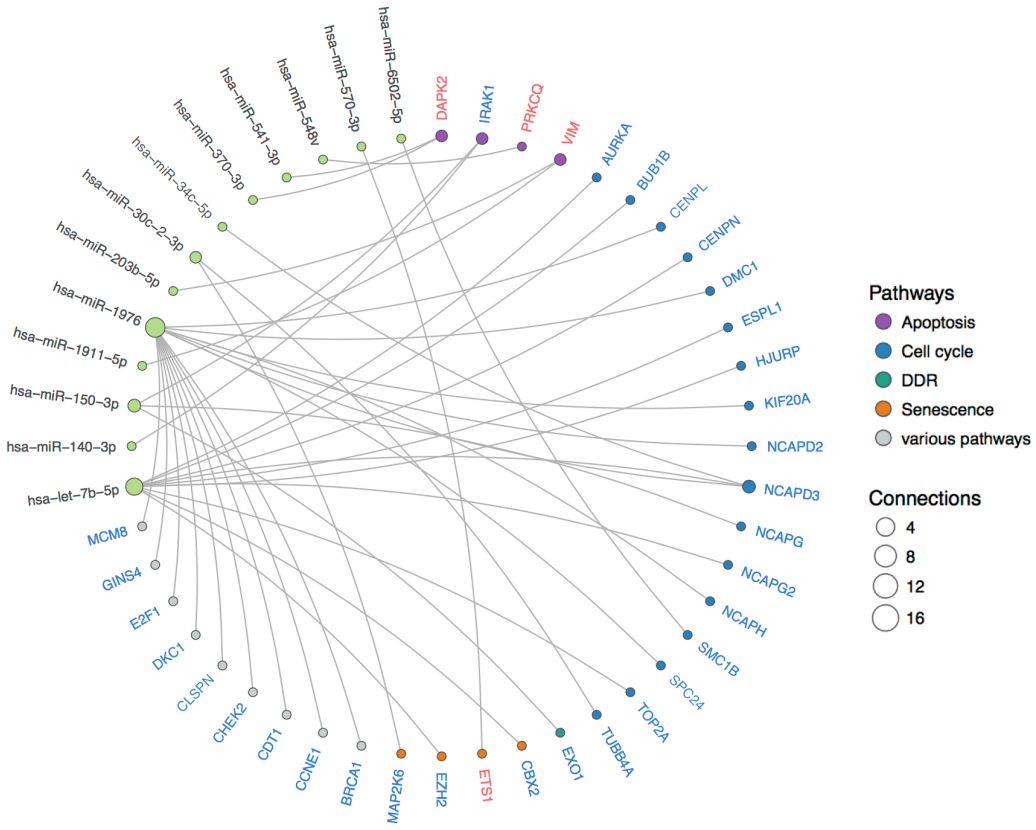


Figura 25. Interacciones exclusivas de pulmón. Interacciones entre miARNs (verde) y genes (azul). La línea continua indica interacciones nuevas, mientras que las líneas discontinuas indican interacciones validadas por otros autores. Se incluye el P-valor ajustado.

Como resultado obtuvimos 11 pares miARN-ARNm relacionados estadísticamente con la supervivencia. Cuatro de ellos resultaron ser significativos para al menos dos tipos tumorales como *HJURP*/let-7c-5p en KIRP y LUAD, *CDC20*/miR-139-3p en KICH y LUAD, el par *KIF20A*/miR-145-5p en HNSC y LUAD, y *CASC5*/miR-139-5p en BRCA y LUAD. Los otros 7 pares se encontraron en un solo tipo de cáncer cada uno (Figura 26) y Tabla Suplementaria 10 del Anexo I).

Posteriormente, se analizaron los pares exclusivos de pulmón usando la misma metodología. De los 40 pares obtenidos (Figura 27), descubrimos 14 en los que los niveles de expresión génica asociaban estadísticamente con la supervivencia del paciente (Figura 27 y Tabla Suplementaria 12 del Anexo I). En particular, 13 fueron obtenidos en LUAD de los cuales 7 están relacionados con el microARN miR-1976 y cuatro con let-7b-5p y sólo uno en LUSC.

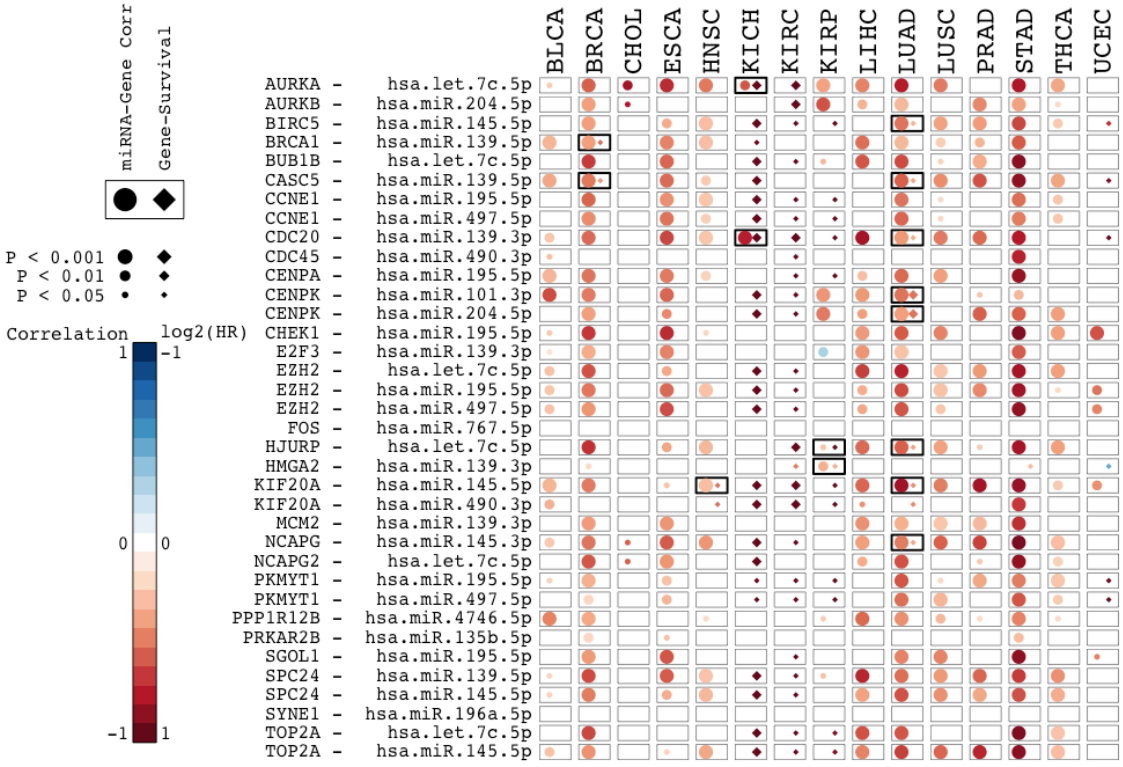


Figura 26. Correlación de expresión y análisis de supervivencia para los 36 pares miARN-ARNm seleccionados. Cada caja resume el resultado de la correlación de expresión y del análisis de supervivencia de para par (filas) en cada tipo tumoral (columna). Los círculos representan la correlación de expresión corregida en base a la metilación y los CNAs. Los diamantes representan el riesgo relativo (HR) obtenido de los modelos de Cox en log2. Valores positivos implican menor tasa de supervivencia. Todos los valores han sido corregidos por FDR y se estima significativo si el FDR<=0.05 tanto en la correlación de expresión como en la supervivencia asociada a los pares.

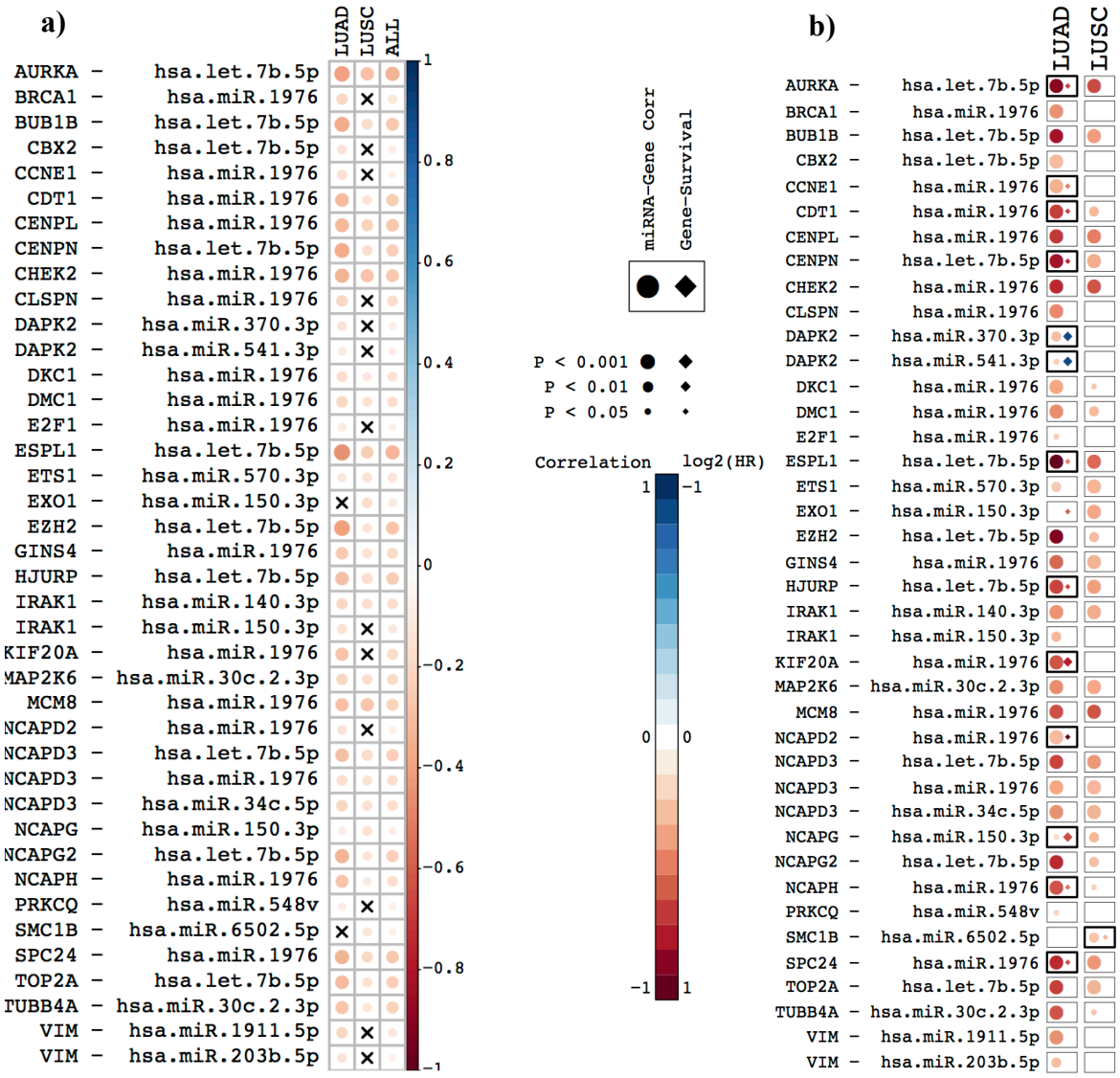


Figura 27. Correlación de expresión entre las interacciones exclusivas de pulmón y su asociación con la supervivencia. A| Correlación analizada mediante una regresión lineal multi-factorial que incluye CNAs, la metilación del gen y la expresión del miARN como variables predictoras de la expresión génica. Este análisis es por cada par (fila) en cada tipo tumoral (columna). El tamaño del circulo es proporcional a la correlación absoluto y el color rojo marca correlación negativa mientras que la azul es positiva. Los modelos incluyen un valor de probabilidad ajustado y aquellos con valores superiores a 0.05 aparecen con una equis. Cuando la expresión del gen o miARN no es detectada, aparece in guión. B| Cada caja resume el resultado de la correlación de expresión y del análisis de supervivencia de para par (filas) en cada tipo tumoral (columna). Los círculos representan la correlación de expresión corregida en base a la metilación y los CNAs. Los diamantes representan el riesgo relativo (HR) obtenido de los modelos de Cox en log2. Valores positivos implican menor tasa de supervivencia. Todos los valores han sido corregidos por FDR y se estima significativo si el FDR<=0.05 tanto en la correlación de expresión como en la supervivencia asociada a los pares.

5. DISCUSIÓN

La identificación de interacciones miARNs-ARNm en cáncer genera un gran interés dado el importante potencial que están asociaciones representan en el diseño de futuros fármacos no cito-tóxicos (Kim, Kang et al. 2016). De esta forma, en la actualidad, diversos fármacos basados en la regulación de microARNs se han llevado a la clínica (Rupaimoole and Slack 2017). La elevada complejidad que acompaña a un estudio de estas características mediante la aplicación de diversas técnicas experimentales, favorece el abordaje a través de análisis computacionales, que permitan la determinación de nuevas dianas terapéuticas, como se ha descrito con éxito en tumores específicos, como el tumor de mama (Ma, Reinhardt et al. 2010) y el tumor de páncreas (Kloosterman, Lagendijk et al. 2007).

La obtención de estas interacciones requiere por un lado poseer de información fiable para la predicción de interacciones y, a su vez, es necesario disponer de un conjunto de genes y microARNs implicados en la tumorogénesis, como los obtenidos mediante el estudio de expresión entre grupos de muestras sanas y tumorales procedentes de pacientes.

De esta forma, en el presente trabajo se ha expuesto por lo tanto: i) el desarrollo de dos herramientas imprescindibles para el estudio de este tipo de interacciones como es miRGate, una base de datos que contiene predicciones con un índice de fiabilidad mayor que el ofrecido por otros repositorios o, miARma-Seq, capaz de analizar muestras de expresión de genes y de microARNs, atendiendo a las distintas necesidades y requisitos de cada tipo de muestra demostrando además, una alta credibilidad. Posteriormente, ambas fueron aplicadas en ii) el análisis computacional de la relación de miARNs/genes en muestras procedentes del Atlas Genómico del Cáncer con el objetivo de obtener interacciones relacionadas con la tumorogénesis y la supervivencia de los pacientes.

Desde el punto de vista del desarrollo, las dos herramientas de manera individual aportan características únicas e importantes con respecto a otros programas, que se discuten a continuación.

5.1. miRGate: base de datos con predicciones de alta fiabilidad.

El objetivo de desarrollar miRGate, es aumentar el número de las asociaciones obtenidas *in silico* por distintos algoritmos de cálculo y que después son satisfactoriamente validadas de forma experimental. En la actualidad, las alternativas existentes proporcionan pares miARN-ARNm que previamente han sido calculados por diferentes algoritmos, así por ejemplo mirGator (Cho, Jang et al. 2013), almacena todas aquellas interacciones facilitadas por algoritmos como son: Pita (Kertesz, Iovino et al. 2007), PicTar (Krek, Grun et al. 2005), Targetscan (Friedman, Farh et al. 2009) y miRanda (Betel, Koppal et al. 2010) en el

momento de su publicación. Tal y como muestra la Tabla 1, este proceso implica el uso tanto de secuencias 3' UTR, como secuencias de miARNs, derivadas de tres versiones diferentes del genoma humano. Por otra parte, MirWalk (Dweep, Gretz et al. 2014) calcula todas las posibles asociaciones entre genes y microARNs usando RNAHybrid (Kruger and Rehmsmeier 2006), pero al igual que el resto de bases de datos, combina estos resultados con las interacciones previamente calculadas por los distintos programas de predicción. En consonancia con lo anteriormente expuesto, al utilizar diferentes construcciones genómicas, tanto la cantidad como el contenido de las secuencias 3' UTR y de las secuencias de los miARNs incluidos, son diferentes.

A su vez, debido a que cuando diversos métodos de predicción emplean una fuente de datos común, existe un incremento considerable en la concordancia de sus resultados (Ritchie, Flamant et al. 2009), hemos diseñado miRGate para utilizar un conjunto de secuencias comunes y actualizadas de los miARNs y las regiones 3' UTR. Por consiguiente, el procesamiento de un conjunto de datos común, como son las secuencias de EnsEMBL y de miRBase por parte de los diferentes métodos de predicción, permite una mejoría en la fiabilidad de las predicciones de entre un 10 y un 21%, según la fiabilidad otorgada a los diversos repositorios de interacciones confirmadas experimentalmente. Estos valores de fiabilidad obtenidos por miRGate, son superiores al del resto de repositorios disponibles.

Asimismo, resulta importante destacar que miRGate, a diferencia de otras herramientas, incluye las secuencias 3' UTR de todas las variantes de cada gen en humano, rata y ratón (incluyendo pseudogenes, transcritos anti-sentido y diversos tipos de genes no codificantes, entre otros). Disponer de un conjunto de secuencias completo, es esencial ya que estas regiones en 3' presentan varios motivos de regulación que controlan la expresión y albergan sitios de unión a miARN, además de poder afectar a la regulación de otros genes. Tal es así, que Poliseno y colaboradores (Poliseno, Salmena et al. 2010) confirmaron esta hipótesis al observar una alteración en la regulación del gen *PTEN* por diversos microARNs debido, a la expresión de un pseudogen con una secuencia 3' UTR de elevada homología a la de *PTEN1*. Trabajar con una sola secuencia 3`-UTR o con aquellas que codifican para proteínas, como implementa la mayoría de las bases de datos, subestima el número de elementos reguladores reales del gen. Esto se debe a que si comparamos dos regiones 3' UTR de diferentes tamaños, aquella de mayor longitud podrá albergar más sitios de unión a miARNs, y por lo tanto, ese ARNm estará expuesto a un nivel más elevado de regulación (Sandberg, Neilson et al. 2008). A su vez, la longitud del 3' UTR

puede afectar no sólo a la estabilidad del ARNm, sino también a su localización, transporte y propiedades de traducción de este ARN (Barrett, Fletcher et al. 2012). Igualmente, existen diversas características dependientes de la secuencia 3' UTR, como son la localización del sitio de unión a lo largo de la región 3' UTR, el contenido de nucleótidos AU alrededor de la zona de complementariedad de secuencia o la cooperación de diversos miARNs a lo largo de la secuencia no traducida.

Finalmente, es interesante destacar que además de los elevados valores de fiabilidad obtenidos por las interacciones almacenadas en nuestra base de datos, diversas relaciones miARN-ARNm propuestas por miRGate y sin ningún tipo de comprobación experimental previa, fueron verificadas posteriormente mediante la aplicación de distintas técnicas de laboratorio. Entre ellas sobresalen, las interacciones en líneas celulares HCC1937 de cáncer de mama (Tanic, Andres et al. 2013), en cáncer de triple negativo de mama (Matamala, Vargas et al. 2016), linfomas de células del manto (Di Lisio, Gomez-Lopez et al. 2010), linfomas de células B (Di Lisio, Martinez et al. 2012), linfomas difusos de células B asociados al virus Epstein-Barr (Martin-Perez, Vargiu et al. 2012), mielomas múltiples híper-diploides (Rio-Machin, Ferreira et al. 2013) y linfomas de Burkitt asociados al virus Epstein-Barr (Ambrosio, Navari et al. 2014).

Por tanto, debido a que en todos los casos miRGate determinó de forma correcta interacciones que después se corroboraron con éxito experimentalmente, podemos afirmar que además de ser una base de datos valiosa para la comunidad científica, también sería una herramienta idónea para llevar a cabo el análisis de las interacciones establecidas entre genes y microARNs diferencialmente expresados en un conjunto de muestras tumorales, obtenidas con la última tecnología de secuenciación (NGS) procedentes de diferentes tipos de tumores.

5.2. miARma-Seq: herramienta para el análisis exhaustivo de muestras de expresión procedentes de técnicas de NGS.

En el análisis de estas muestras de transcriptómica procedentes de técnicas de NGS, se utilizan diferentes herramientas que procesan eficazmente este tipo de datos (Figura 10). Sin embargo, la mayoría de estas herramientas requieren elevados conocimientos informáticos para su uso. En este sentido, la comunidad científica que desarrolla programas para el estudio de muestras de NGS, ha facilitado de forma notable el manejo de estas herramientas. Actualmente, la interoperabilidad del software no supone una limitación, ya que permiten la instalación local en distintos sistemas operativos. Una estrategia

frecuentemente aplicada para simplificar el uso de estos programas, es el desarrollo de versiones accesibles desde páginas web, aunque conlleve importantes restricciones. Concretamente, el conjunto de los sistemas basados en la web como son Galaxy o RAP (D'Antonio, D'Onorio De Meo et al. 2015), presentan problemas intrínsecamente relacionados con el tráfico de datos a través de internet (es decir, límites de ancho de banda, políticas de privacidad de datos que afectan a los usuarios, y/o saturación de colas de ejecución de procesos, entre otros). Además, muestran limitaciones en el número total de muestras a analizar o en el espacio disponible para almacenar los datos necesarios, como es el caso de RAP. Incluso, en los sistemas que permiten su instalación de forma local para reducir estos inconvenientes, precisan de un alto conocimiento en la instalación de programas bajo entornos Unix, como sucede con Galaxy.

Por consiguiente, la principal dificultad en el uso de estos métodos de uso local es la instalación del propio programa y de las dependencias impuestas para su correcto funcionamiento, que por lo general, requieren de gran competencia en la administración de sistemas operativos. Como resultado, estas importantes limitaciones provocan una utilización minoritaria de las herramientas de análisis por parte de investigadores no experimentados.

Dado que el objetivo principal del presente trabajo, demanda el estudio de un elevado número de muestras, resultaba imprescindible el desarrollo de una nueva herramienta capaz de ejecutar un análisis completo, tanto de muestras de miRNA-Seq como de muestras de RNA-Seq, y que a su vez, se pudiera manejar de forma automática, sin solicitar la intervención del usuario en las distintas etapas del análisis. Por esta razón, creamos miARma-Seq, un programa rápido y eficaz en el procesamiento de datos de RNA-Seq y miRNA-Seq, que permite identificar ARNm, miARNs y ARNs circulares y, que resuelve las principales limitaciones con las que se encuentran los investigadores al analizar los datos de secuenciación, como son: i) capacidad de análisis integrado para datos de miARNs y datos de expresión de ARNm; ii) simplicidad en la instalación, eliminando la necesidad de instalar programas y dependencias externas que dificultan su uso; iii) permitir el análisis de muestras procedentes de cualquier organismo con un genoma de referencia; iv) flexibilidad en la adaptación a cualquier diseño experimental; v) alta capacidad de cálculo, permitiendo realizar el análisis de muestras tanto en un equipo estándar como en sistemas de computación de alto rendimiento, aprovechando su entorno de paralelización; vi) fiabilidad, debida a que miARma-Seq integra diversos algoritmos de análisis ampliamente aceptados en la comunidad científica; vii) extensa cobertura, ya que no sólo incluye el

análisis de muestras de miRNA-Seq y RNA-Seq para identificar miARNs, ARNm o circARNs, sino que también, permite ejecutar análisis de expresión diferencial, análisis de enriquecimiento funcional o predecir miARNs *de novo*, es decir no descritos en bases de datos como miRBase (Kozomara and Griffiths-Jones 2014).

En el desarrollo de un programa de estas características, además de incorporar funciones que otras herramientas no ofrecen, es imprescindible comprobar los resultados obtenidos como medida de fiabilidad. En este sentido, hemos contrastado los resultados generados por miARma-Seq procedentes de diversos conjuntos de datos publicados, y comparado estos resultados con los datos obtenidos por otros autores utilizando herramientas diferentes a miARma-Seq. Tal y como se explica en el apartado de resultados, la correlación tanto en los valores de expresión, como en los datos de expresión diferencial de microARNs y genes, demuestra la elevada fiabilidad y valor de esta nueva herramienta desarrollada, para la comunidad de científica.

En conclusión, miARma-Seq es una herramienta notablemente flexible y eficaz, que permite el análisis de datos de transcriptómica (miARN, ARNm y circARNs). Además, ofrece importantes posibilidades como son: la identificación de entidades diferencialmente expresadas, predicción de miARN-ARNm o análisis funcionales (Figura 10). miARma-Seq se caracteriza por su simplicidad, similar a la de una herramienta web, excluyendo las limitaciones relacionadas con la privacidad de las muestras de pacientes o el tráfico de red de estas aplicaciones. Por otra parte, al tratarse de una herramienta de instalación local, hace posible el análisis de una gran cantidad de muestras de forma simultánea debido a su diseño paralelizado (multi-hilo), pero eliminando las dependencias externas que requieren las herramientas de uso local. Todas estas propiedades favorecen que miARma-Seq sea aplicada por una extensa variedad de usuarios, desde aquellos que presentan escasa experiencia en las áreas de programación e informática, tanto en instalación como en uso, hasta usuarios avanzados que dominan la línea de comandos.

5.3. Interacciones miARNs-ARNm conservadas en cáncer.

Una vez desarrolladas las metodologías necesarias para abordar parte de los objetivos específicos previamente expuestos, nos planteamos llevar a cabo el objetivo principal mediante el uso combinado de estas herramientas y así estudiar las redes de regulación entre microARNs y genes implicados en la tumorogénesis, procedentes de distintos tipos tumorales.

La relación entre genes y cáncer se remonta a los años 50, cuando la primera teoría de los oncogenes fue expuesta por el físico Niels Henrik Arley. El primer oncogen fue descubierto sin embargo en 1976, cuando Dominique Stehelin, J. Michael Bishop y Harold E. Varmus procedentes de la Universidad de California, demostraron la teoría de los oncogenes, lo que les valió el premio Nobel de medicina en 1989. Sin embargo, la relación entre los microARNs y el cáncer es mucho más actual. En un principio fueron relacionados con el desarrollo y la diferenciación, dado su papel en Caenorhabditis elegans (Lee, Feinbaum et al. 1993) donde fueron descubiertos. Esto promovió el desarrollo de nuevas líneas de investigación que permitieron siete años después, identificar el primer miARN conocido en mamíferos: let-7. Tan solo dos años después, en 2002, se identificaron los primeros microARNs cuya desregulación estaba implicada en cáncer (Calin, Dumitru et al. 2002).

En la actualidad varias líneas de investigación han elucidado que los miARNs se expresan diferencialmente en las células tumorales, donde crean un patrón de expresión único (Lu, Getz et al. 2005) y su desregulación está hoy en día reconocida como una propiedad común en los distintos tipos de cáncer. Como resultado, estos descubrimientos han favorecido el diseño de diversos fármacos, concebidos para frenar el crecimiento del cáncer en base a este fenómeno (Rupaimoole and Slack 2017).

Sin embargo, la mayoría de los estudios efectuados hasta la fecha, se han basado en el análisis de tumores individuales y apenas se han llevado a cabo análisis sistemáticos, para inferir interacciones miARN-ARNm procedentes de distintos tipos de tumores. En esta línea de investigación, un método denominado REC (del inglés *Recurrent Score*) (Jacobsen, Silber et al. 2013) fue desarrollado para evaluar la recurrencia de pares miARN-ARNm en distintos tipos de tumores. Este método, al igual que PanMira (Li and Zhang 2014), utiliza datos de expresión procesados por terceros, combinando datos de microarrays y de RNA-Seq sólo de muestras tumorales, sin incluir muestras control de individuos sanos. Por el contrario, la mayoría de trabajos publicados basados en el consorcio del atlas genómico del cáncer (TCGA), se basan en estudios de expresión diferencial entre muestras de pacientes sanos/enfermos.

Igualmente, en esta tesis doctoral se han incorporado muestras control de tejido sano, lo que permite identificar asociaciones miARN-ARNm formadas tanto por genes como por microARNs diferencialmente expresados entre ambos tipos de muestras, descartando de esta forma, interacciones no relacionadas con cáncer. Por lo tanto, este trabajo ha abordado el análisis en profundidad, de 15 de los tipos tumorales más frecuentes. Concretamente, se

ha realizado un estudio exhaustivo de los datos brutos de 18.605 muestras de RNASeq y miRNASeq, siguiendo los protocolos de análisis recomendados (Anders, McCarthy et al. 2013) y aplicando filtros restrictivos para reducir la aparición de falsos positivos. Además, hemos enfocamos el estudio hacia las vías relevantes en cáncer (Hanahan and Weinberg 2011), ya que éstas son más susceptibles a la identificación de genes y sus microARNs reguladores, candidatos para fines terapéuticos.

Tras este análisis, se obtuvo para cada tipo de tumor un listado de genes y miARNs diferencialmente expresados, concordantes con los resultados publicados previamente por otros autores. Específicamente, la mayoría de los genes resultantes que aparecen sobre-expresados representan oncogenes, que definen un mal pronóstico en la progresión tumoral, y un elevado porcentaje de los genes inhibidos resultaron ser genes supresores de tumores. De esta forma, obtuvimos que el oncogen *HMGA2* aparecía especialmente sobre-expresión en cholangiocarcinoma (CHOL) (Lee, Wu et al. 2014), tumores de cuello y cabeza (HNSC) (Yamazaki, Mori et al. 2013), tumores de pulmón (LUAD y LUSC) (Meyer, Loeschke et al. 2007) y carcinoma de tiroides (THCA) (Belge, Meyer et al. 2008). En todos estos tumores, se asoció la sobre-expresión de este gen con un mal pronóstico y ligado a procesos invasivos de metástasis (Morishita, Zaidi et al. 2013), además de estar relacionada con la transformación de células neoplásicas (Fusco and Fedele 2007). De igual forma, el oncogen *BIRC7* (también conocido como Livin) aparece entre los genes más sobre-expresados en carcinoma de estómago (STAD) (Wang, Ding et al. 2010), de esófago (ESCA) (Zhang, Tang et al. 2014) y en los tumores de riñón (KICH, KIRC y KIRP) (Crnkovic-Mertens, Wagener et al. 2007). La expresión exacerbada de *BIRC7* está relacionada con la formación y proliferación tumoral (Yan 2011). Por el contrario, también encontramos ejemplos de genes que aparecen bajo una alta represión, entre ellos destacamos a *HSPA2* en los tres tipo de tumores de riñón. La inhibición del nivel de expresión de este gen, está asociada con el desarrollo del cáncer y la invasión tumoral (Singh and Suri 2014). *FOS*, inicialmente reconocido como oncogen, aparece como el gen más reprimido en el tumor de hígado. Esta represión fue estudiada por Teng y colaboradores en diferentes tipos celulares y comprobaron que la disminución de la expresión implicaba en diversos casos, una disminución de la apoptosis, más propiamente vinculada a un gen supresor de tumores (Teng 2000). Posteriormente estos resultados se confirmaron en el tumor de hígado (Mikula, Gotzmann et al. 2003), y en el tumor de ovario (Mahner, Baasch et al. 2008). Interesantemente, *FOS* aparece como el tercer gen mas reprimido en nuestro estudio de tumor de endometrio.

El papel de los miARNs también es fundamental en los tumores individuales, por este motivo se analizó una elevada cantidad de muestras de distintos tipos de tumores para obtener un listado de microARNs desregulados e implicados en el fenotipo tumoral. Según nuestros resultados, podríamos destacar que diversos oncomires mostraban una alta expresión sólo en las muestras tumorales, entre ellos destacan miR-184 y miR-891a del clúster miR-888, siendo los miARNs más alterados en el cáncer de próstata (PRAD). Esto ha sido relacionado por otros investigadores con el crecimiento y mal pronóstico del cáncer de próstata (Lewis, Lance et al. 2014). Por el contrario, también se obtuvieron miARNs fuertemente inhibidos, identificados como microARNs supresores de tumores. Concretamente, anti-oncomires como miR-139, miR-143/145 y miR-195, aparecen altamente reprimidos en los tumores de mama (BRCA), vejiga (BLCA) y estómago (STAD), como a su vez determinaron otros autores (Wu, Xue et al. , Das and Pillai 2015, Yonemori, Seki et al. 2016). Igualmente, en carcinoma hepatocelular (LIHC), destaca la alta inhibición de miR-424, relacionada con un incremento en la proliferación celular, la migración e invasión tumoral (Yu, Ding et al. 2014). Por último, obtuvimos que el supresor tumoral miR-202, aparecía inhibido en los tumores uterinos de endometrio, fenómeno que se ha relacionado con el desarrollo de este tipo de carcinomas (Hiroki, Akahira et al. 2010).
La elevada consistencia existente entre los resultados obtenidos en nuestro estudio de tumores individuales con numerosas publicaciones anteriores, nos permitió llevar a cabo un estudio de tipo pan-cáncer y así, determinar interacciones establecidas entre genes y microARNs desregulados, propias de la mayoría de los distintos tipos de tumores incluidos, como de forma exclusiva en tumores procedentes de un mismo órgano. En este nuevo análisis, destacando en primer lugar los genes resultantes (Figura 20), subrayamos que el gen *SPC24* aparece significativamente sobre-expresado en la totalidad de los tumores estudiados. Además, el aumento en la expresión de este gen en muestras tumorales ha sido asociado con un mal pronóstico tanto en cáncer colorrectal como en cáncer gástrico (Kaneko, Miura et al. 2009). Otros ejemplos son los genes *E2F1* y Claspin (*CLSPN*), desregulados en 14 y 12 tipos de tumores, y reconocidos por sus papeles oncogénicos y supresores de tumores (Pierce, Schneider-Broussard et al. 1999), respectivamente. Por el contrario, el gen *SYNE1* muestra una inhibición relevante en la mayoría de tumores de nuestro estudio (Figura 20), y dada su función esencial en la migración del centrosoma, y la alteración en su nivel de expresión, se ha relacionado con el desarrollo de distintos tipos de tumores como glioblastomas (Serao, Delfino et al. 2011) o tumores de ovario (Doherty, Rossing et al. 2010), entre otros.

Interesantemente, identificamos diversos microARNs que de forma constante, aparecían estadísticamente desregulados en casi la totalidad de los tumores analizados. Entre ellos, destaca el miR-183 del "Onco-clúster" miR-183 (Dambal, Shah et al. 2015), cuya expresión se encuentra significativamente alterada en 14 de los 15 tipos de tumores estudiados, reforzando su papel pro-tumoral, como ya se había demostrado en meduloblastomas (Weeraratne, Amani et al. 2012). Por otra parte, relacionado con la progresión tumoral, encontramos a miR-4746, el único microARN que aparece sobre-expresado en el conjunto de los tumores incluidos en este trabajo, confirmando publicaciones previas (Zhou, Zhou et al. 2015). A su vez, otros microARNs que desempeñen el papel de supresores tumorales como: miR-145, miR-139 y miR-195, se muestran en consecuencia, considerablemente inhibidos según nuestros datos.

Por tanto, los resultados globales obtenidos en el análisis de muestras de expresión de microARNs, como de genes, son ratificados por numerosos estudios publicados, quedando demostrada de esta manera, la elevada fiabilidad y eficacia que caracteriza a nuestro análisis y las herramientas aplicadas en la determinación de interacciones establecidas entre miARNs y genes diferencialmente expresados.

Una vez obtenido el listado de genes y microARNs diferencialmente expresados, mediante el uso de miRGate, se obtuvieron 41 interacciones conservadas, donde ambos elementos aparecían desregulados con valores de expresión opuestos en al menos 8 tipos de tumores diferentes (Figura 22 y Tabla Suplementaria 6 del Anexo I). Dado que es conocido que tanto la metilación como la alteración en el número de copias afectan la expresión de los genes, especialmente en cáncer, estas interacciones fueron filtradas para tener en cuenta estos fenómenos. De esta forma, obtuvimos un total de 36 asociaciones, de las cuales 25, no han sido descritas con anterioridad por ningún autor, aunque individualmente la mayoría de los genes y miARNs implicados, están relacionados de alguna manera con el desarrollo del cáncer. Concretamente, en el caso de la interacción *SCP24*/miR-139, en la que, la implicación de ambos componentes por separado está bien establecida (*SPC24* (Thiru, Kern et al. 2014) y miR-139-5p (Guo, Miao et al. 2009)) pero, no existe ninguna evidencia experimental que haya corroborado una relación de regulación entre ambos.

Asimismo, 17 de las 36 interacciones descubiertas, destacaron debido a su elevada especificidad, dado que los interactores que definen esa asociación, aparecen estadísticamente asociados con un número muy bajo de interactores alternativos. Este tipo de asociaciones son especialmente relevantes en la clínica, dado que podrían representar futuras dianas terapéuticas para combatir el cáncer, reduciendo los efectos secundarios que

aparecen con frecuencia en los fármacos sintetizados a partir de oligonucleótidos anti-microARNs (AMOs).

Posteriormente, comparamos las interacciones miARN-ARNm entre los distintos tipos tumorales. Así observamos que, aquellos tumores procedentes de un mismo órgano exhiben interacciones no identificadas en el resto de tumores (Figura 25). En particular, en los tumores originarios del pulmón (LUAD y LUSC) encontramos una alta cantidad de interacciones mediadas por miR-1976 y let-7b-5p. Estos parecen actuar como reguladores claves de genes relacionados con el ciclo celular y la apoptosis, sugiriendo la posibilidad de convertirse en importantes biomarcadores de cáncer de pulmón. Tal es así, que el papel de miR-1976 como supresor tumoral en cáncer de pulmón, se ha descrito recientemente en relación al oncogen *PLCE1* (Chen, Hu et al. 2016).

Existen muchas factores que afectan al pronóstico del cáncer. La supervivencia depende en última instancia de una red muy compleja de mecanismos entre los que se incluyen la agresividad del tumor, la posibilidad de extirpación, el tratamiento primario rutinario, la resistencia a los fármacos, la edad del paciente, su estado físico, comorbilidad, etc. Todos estos factores pueden variar según el tipo de tumor y así afectar notablemente al pronostico. Durante años, los médicos han empleado una serie de características clínico-patológicas que han demostrado ser muy útiles para estimar la supervivencia del paciente con cáncer. En particular, el estadío tumoral y el grado de diferenciación se han utilizado como factores pronósticos en el diagnóstico y tratamiento de la mayoría de los tipos de cáncer. Por ejemplo, los pacientes con cánceres de estadío III o IV presentan células malignas que desarrollan mecanismos para promover la invasión a los ganglios linfáticos o a órganos colindantes, y por consiguiente, presentan un pronóstico mucho peor que los pacientes con tumores en estadio I-II. La edad y el sexo del paciente, también se emplean típicamente como marcadores predictivos de supervivencia. En los últimos años, los marcadores moleculares de la supervivencia del paciente o de la respuesta al tratamiento, se han incorporado constantemente en el ámbito clínico (ASCO, https://www.asco.org/practice-guidelines y, http://www.esmo.org/Guidelines/Guidelines-News, ESMO).

Por lo tanto, debido a la enorme complejidad y a la plausible interacción de muy diversos factores implicados en la supervivencia, no se debe esperar que un único elemento (como un gen o un microARN) produzca un gran efecto sobre la supervivencia por sí mismo, especialmente a través de tipos diferentes de tumores. No obstante, hemos obtenido interacciones asociadas a la supervivencia, algunas de los cuales ya han sido extensamente

estudiadas en muchos tipos de tumores, como por ejemplo la asociada a la desregulación de *AURKA*(Siggelkow, Boehm et al. 2012, Goos, Coupe et al. 2013, Zhang, Li et al. 2015) o *BRCA1* (Lambie, Miremadi et al. 2003, Lesnock, Darcy et al. 2013). Sin embargo, también descubrimos genes no descritos en la literatura y asociados a la supervivencia, como *CASC5*. Este gen se describió previamente en el cáncer de pulmón (Takimoto, Wei et al. 2002), pero no asociado a cáncer de mama, como hemos encontrado en este estudio. También, merece la pena mencionar que su miARN regulador (miR-139-p) también ha sido asociado a un mal pronóstico en el cáncer de pulmón (Sun, Sang et al. 2015), pero nuevamente no en BRCA. Por lo tanto, nuestro estudio ha revelado una nueva interacción miARN-gen asociada la supervivencia en el cáncer de mama, respaldada por hallazgos anteriores en el cáncer de pulmón. Igualmente, obtuvimos una relación estadísticamente significativa entre el gen *BIRC5* y el adenocarcinoma de pulmón (LUAD) no formulada hasta la fecha, aunque los niveles de expresión de este gen mostraron afectar la supervivencia en mama (Lv, Yu et al. 2010), vejiga (Akhtar, Gallagher et al. 2006) y sarcoma (Taubert, Kappler et al. 2005). Sin embargo, su microARN asociado (miR-145-5p) ha sido vinculado con un pronóstico deficiente en tumores gástricos(Zhang, Wen et al. 2016) y renales (Slaby, Redova et al. 2012). A su vez, *CDC20* ha sido descrito como factor asociado a un mal pronóstico en colon (Wu, Hu et al. 2013) y mama (Karra, Repo et al. 2014), en este estudio ha sido identificado como relevante en KICH y LUAD. Finalmente *NCAPG* y *KIF20A* son genes asociados a pronóstico en gliomas(Duan, Huang et al. 2016, Liang, Hsieh et al. 2016) y en nuestro estudio en LUAD. También identificamos algunas nuevas asociaciones en tumores de pulmón (LUAD), por ejemplo *CENPK* no se ha asociado previamente con este tipo de cáncer, a pesar de que el gen ya había sido asociado a cáncer de ovario (Lee, Huang et al. 2015). Hasta la fecha, los genes *CDT1, CENPN, DAPK2, HJURP, KIF20A, NACPD2, NCAPG, NCAPH, SMC1B* y *SPC24* no se han asociado a la supervivencia en el cáncer de pulmón, aunque la mayoría de sus microARNs si han sido vinculados. Por ejemplo, los microARNs claves en la regulación de las interacciones exclusivas de pulmón hsa-miR-1976 y hsa-let-7b-5p, ya habían sido asociados a una mala supervivencia en cáncer de pulmón (Jusufovic, Rijavec et al. 2012, Chen, Hu et al. 2016) y en otros tumores como el de mama (Ma, Li et al. 2014), próstata (Wang, Xu et al. 2016) y el gástrico (Kang, Tong et al. 2014). En conclusión, nuestro análisis ha revelado nuevas interacciones miARN-gen correlacionadas con el pronóstico de paciente, apoyado por hallazgos anteriores en distintos tipos de tumores.

En resumen, además de obtener unos resultados comparables a otros autores en los análisis realizados a los 15 tumores individuales, nuestro análisis global mediante la implementación de una metodología estadísticamente confiable, proporcionó nuevas interacciones miARN-ARNm vinculadas a las vías relevantes en el cáncer procedentes de diferentes tipos de tumores. Además, la posterior identificación de interacciones asociadas significativamente con la supervivencia, abre nuevas vías para la investigación. Estos análisis exploratorios, seguidos por una futura validación experimental, podría ayudar a esclarecer el papel de un gran conjunto de miARNs en la tumorogénesis. Como ejemplo, la asociación descubierta en el tumor de mama para el gen *BRCA1* integrado junto con un análisis de mutaciones somáticas podría explicar casos de tumores donde la tumorigenesis no se pueda afirmar por mutaciones en este gen.

Estos resultados podrían confirmar la validez de nuestro estudio orientado a la identificación de interacciones miARN-ARNm que permitan el diseño de nuevos fármacos que minimicen los efectos secundarios y así además, mejorar los tratamientos ya existentes.

5.4. Otros aspectos.

En los análisis realizados en este tasis doctoral, se ha tenido en cuenta como la metilación de las islas CpGs, la alteración en el número de copias y la regulación llevada por miARNs afecta a la expresión génica. En el caso de la regulación llevada a cabo por los microARNs, se ha tenido en cuenta el procedimiento de regulación negativa basado en la unión de la secuencia semilla del miARN con una región complementaria en el 3'-UTR, dado que este mecanismo es el más conocido y estudiado. Sin embargo, se han descrito mecanismos adicionales basados en la unión de microARNs a sitios no canónicos (Lytle, Yario et al. 2007, Helwak, Kudla et al. 2013). Asimismo, se ha demostrado que la gran mayoría de las interacciones formadas por los microARNs se establecen entre las regiones semilla y los 3'-UTR de las isoformas génicas, y en base a ello, nuestro diseño experimental se ha basado en la correlación inversa de expresión, dado que la hipótesis subyacente a este estudio es que los miARNs inhiben sus genes diana a través de su 3 'UTR. Por otra parte, estudios previos en mamíferos, indican que los microARNs tienden predominantemente a disminuir los niveles de los mARN a los que se unen (Guo, Ingolia et al. 2010), donde la desestabilización del mARN representa más del 84% de la disminución de la producción de proteínas.

Dada la complejidad adicional de los mecanismos de unión no canónicos, muchos aun no elucidados, y la limitación técnica de los programas de predicción existentes y basados

principalmente en el estudio de regiones 3`-UTR, es posible que información importante en la regulación basada en uniones no canónicas de genes/microARNs no haya sido identificada.

5.5. Perspectivas.

Como se comentó en la introducción, la tumorogénesis se debe a diversos factores genéticos y epigenéticos, muchos de ellos aún desconocidos. En este trabajo se ha intentando profundizar en el mecanismo relacionado con los cambios de expresión de los genes y por lo tanto de las proteínas que codifican, dado que la sobre-expresión de genes (oncogenes) con funciones relacionadas con la proliferación, favorecen el desarrollo de tumores y por el contrario, la inhibición de expresión de genes codificantes para proteínas implicadas en barreras anti-proliferativas (como senescencia o apoptosis), igualmente favorecen la tumorogénesis. Pero de similar manera, las mutaciones puede generar proteínas no funcionales que a su vez favorezcan el crecimiento tumoral. En base a lo explicado, nos gustaría integrar los datos de mutaciones en los genes identificados.

A su vez creemos interesante el estudio de regulación de genes en base a los diversos microARNs que puedan regularle de forma aditiva, así como aquellos ARNs reguladores a su vez de los microARNs, como ARNs circulares o genes no codificantes de largo tamaño (lncARN).

Por ultimo, sería de gran interés estudiar la relación entre los cambios de expresión génica y los cambios en los niveles cuantificados de las proteínas que codifican así como el efecto de las mutaciones. En ese sentido el atlas genómico del cáncer está haciendo un gran esfuerzo para incluir datos proteómicos de pacientes diagnosticados con diversos tipos tumorales. De esta forma, actualmente están disponibles datos de espectrometría de masas procedente de 95 pacientes de tumores colorrectal (Zhang, Wang et al. 2014), 105 pacientes de tumor de mama (Mertins, Mani et al. 2016) y 174 pacientes de tumor de ovario (Mertins, Mani et al. 2016).

6. CONCLUSIONES

En este trabajo, se han extraído las siguientes conclusiones:

1. Las interacciones miARN-ARNm obtenidas mediante miRGate, al compararse con un conjunto de datos validados experimentalmente, mostraron un alto nivel de fiabilidad. Aquellos pares no validados hasta la fecha, pero predichos con una buena coincidencia genómica, fueron posteriormente validados en diversas colaboraciones.

2. La metodología y las herramientas incluidas en miARma-Seq, permiten la obtención de miARNs y genes estadísticamente desregulados en muestras de transcriptómica de tumores individuales, en consonancia con lo expuesto por otros autores.

3. El estudio global de genes diferencialmente expresados en un conjunto de 15 tumores distintos, permitió determinar genes sobre-expresados con clara identidad oncogénica (*CDT1*, *DTL*, *SPC24*). Por el contrario, los genes reprimidos que obtuvimos en la mayoría de los tumores, aparecen identificados con función supresora de tumores (*SYNE1*, *FOS*).

4. El análisis integrativo del grupo de los diferentes tumores, confirmó la existencia de oncomires diferencialmente sobre-expresados en la mayoría de los tumores, como miR-196a, miR-181b o miR-767, así como una mayoría de miARNs reprimidos como los anti-oncomires miR-let-7c, miR-145 o miR-139.

5. A través de la integración de genes y miARNs diferencialmente expresados y teniendo en cuenta la participación de la metilación y la alteración en el número de copias génicas, hemos establecido 36 interacciones relacionadas con el fenotipo tumoral, 17 de ellas de alta especificidad. De estas interacciones, al menos hay 25 de ellas no publicadas actualmente por otros autores y que afectan principalmente a funciones esenciales del ciclo celular y la apoptosis.

6. El estudio de interacciones entre tumores procedentes de un mismo órgano, nos ha permitido obtener 40 interacciones exclusivas entre los tumores de pulmón. En ellas observamos que tanto el miR-1976 como el let-7b-5p regulan un elevado número de genes involucrados en rutas relacionadas con el ciclo celular y la apoptosis.

7. El análisis posterior de las 36 interacciones conservadas y las 40 asociaciones exclusivas de pulmón obtenidas, nos ha permitido obtener 11 y 14 interacciones respectivamente, implicadas en la supervivencia de los pacientes.

7. REFERENCIAS

Adams, J. M. and S. Cory (2007). "The Bcl-2 apoptotic switch in cancer development and therapy." Oncogene **26**(9): 1324-1337.

Akhtar, M., L. Gallagher and S. Rohan (2006). "Survivin: role in diagnosis, prognosis, and treatment of bladder cancer." Adv Anat Pathol **13**(3): 122-126.

Albertella, M. R., A. Lau and M. J. O'Connor (2005). "The overexpression of specialized DNA polymerases in cancer." DNA Repair (Amst) **4**(5): 583-593.

Ambrosio, M. R., M. Navari, L. Di Lisio, E. A. Leon, A. Onnis, S. Gazaneo, L. Mundo, C. Ulivieri, G. Gomez, S. Lazzi, M. A. Piris, L. Leoncini and G. De Falco (2014). "The Epstein Barr-encoded BART-6-3p microRNA affects regulation of cell growth and immuno response in Burkitt lymphoma." Infect Agent Cancer **9**: 12.

An, O., G. M. Dall'Olio, T. P. Mourikis and F. D. Ciccarelli (2016). "NCG 5.0: updates of a manually curated repository of cancer genes and associated properties from cancer mutational screenings." Nucleic Acids Res **44**(D1): D992-999.

Anders, S. (2010). "FastQC: a quality control tool for high throughput sequence data." Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Anders, S., D. J. McCarthy, Y. Chen, M. Okoniewski, G. K. Smyth, W. Huber and M. D. Robinson (2013). "Count-based differential expression analysis of RNA sequencing data using R and Bioconductor." Nat Protoc **8**(9): 1765-1786.

Andres Leon, E., G. Gomez-Lopez and D. G. Pisano (2017). "Prediction of miRNA–mRNA Interactions Using miRGate." Methods Mol Biol.

Andres-Leon, E., I. Cases, S. Alonso and A. M. Rojas (2017). "Novel miRNA-mRNA interactions conserved in essential cancer pathways." Sci Rep.

Andres-Leon, E., I. Cases, A. Arcas and A. M. Rojas (2016). "DDRprot: a database of DNA damage response-related proteins." Database (Oxford) **2016**.

Andres-Leon, E., D. Gonzalez Pena, G. Gomez-Lopez and D. G. Pisano (2015). "miRGate: a curated database of human, mouse and rat miRNA-mRNA targets." Database (Oxford) **2015**: bav035.

Andres-Leon, E., R. Nunez-Torres and A. M. Rojas (2016). "miARma-Seq: a comprehensive tool for miRNA, mRNA and circRNA analysis." Sci Rep **6**: 25749.

Arcas, A., O. Fernandez-Capetillo, I. Cases and A. M. Rojas (2014). "Emergence and evolutionary analysis of the human DDR network: implications in comparative genomics and downstream analyses." Mol Biol Evol **31**(4): 940-961.

Baek, D., J. Villen, C. Shin, F. D. Camargo, S. P. Gygi and D. P. Bartel (2008). "The impact of microRNAs on protein output." Nature **455**(7209): 64-71.

Baeriswyl, V. and G. Christofori (2009). "The angiogenic switch in carcinogenesis." Semin Cancer Biol **19**(5): 329-337.

Bandyopadhyay, D., N. A. Okan, E. Bales, L. Nascimento, P. A. Cole and E. E. Medrano (2002). "Down-regulation of p300/CBP histone acetyltransferase activates a senescence checkpoint in human melanocytes." Cancer Res **62**(21): 6231-6239.

Barrallo-Gimeno, A. and M. A. Nieto (2005). "The Snail genes as inducers of cell movement and survival: implications in development and cancer." Development **132**(14): 3151-3161.

Barrett, L. W., S. Fletcher and S. D. Wilton (2012). "Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements." Cell Mol Life Sci **69**(21): 3613-3634.

Bartek, J., C. Lukas and J. Lukas (2004). "Checking on DNA damage in S phase." Nat Rev Mol Cell Biol **5**(10): 792-804.

Bartek, J. and J. Lukas (2001). "Mammalian G1- and S-phase checkpoints in response to DNA damage." Curr Opin Cell Biol **13**(6): 738-747.

Bartel, D. P. (2004). "MicroRNAs: genomics, biogenesis, mechanism, and function." Cell **116**(2): 281-297.

Belge, G., A. Meyer, M. Klemke, K. Burchardt, C. Stern, W. Wosniok, S. Loeschke and J. Bullerdiek (2008). "Upregulation of HMGA2 in thyroid carcinomas: a novel molecular marker to distinguish between benign and malignant follicular neoplasias." Genes Chromosomes Cancer **47**(1): 56-63.

Benjamini Y, H. Y. (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." Journal of the Royal Statistical Society **57**: 11.

Bernstein, E., A. A. Caudy, S. M. Hammond and G. J. Hannon (2001). "Role for a bidentate ribonuclease in the initiation step of RNA interference." Nature **409**(6818): 363-366.

Betel, D., A. Koppal, P. Agius, C. Sander and C. Leslie (2010). "Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites." Genome Biol **11**(8): R90.

Bhowmick, N. A., E. G. Neilson and H. L. Moses (2004). "Stromal fibroblasts in cancer initiation and progression." Nature **432**(7015): 332-337.

Bisognin, A., G. Sales, A. Coppe, S. Bortoluzzi and C. Romualdi (2012). "MAGIA(2): from miRNA and genes expression data integrative analysis to microRNA-transcription factor mixed regulatory circuits (2012 update)." Nucleic Acids Res **40**(Web Server issue): W13-21.

Blasco, M. A. (2005). "Telomeres and human disease: ageing, cancer and beyond." Nat Rev Genet **6**(8): 611-622.

Bohnsack, M. T., K. Czaplinski and D. Gorlich (2004). "Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs." RNA **10**(2): 185-191.

Braig, M. and C. A. Schmitt (2006). "Oncogene-induced senescence: putting the brakes on tumor development." Cancer Res **66**(6): 2881-2884.

Cadet, J., J. L. Ravanat, M. TavernaPorro, H. Menoni and D. Angelov (2012). "Oxidatively generated complex DNA damage: tandem and clustered lesions." Cancer Lett **327**(1-2): 5-15.

Calin, G. A., C. D. Dumitru, M. Shimizu, R. Bichi, S. Zupo, E. Noch, H. Aldler, S. Rattan, M. Keating, K. Rai, L. Rassenti, T. Kipps, M. Negrini, F. Bullrich and C. M. Croce (2002). "Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia." Proc Natl Acad Sci U S A **99**(24): 15524-15529.

Campisi, J. (2001). "Cellular senescence as a tumor-suppressor mechanism." Trends Cell Biol **11**(11): S27-31.

Camps, C., H. K. Saini, D. R. Mole, H. Choudhry, M. Reczko, J. A. Guerra-Assuncao, Y. M. Tian, F. M. Buffa, A. L. Harris, A. G. Hatzigeorgiou, A. J. Enright and J. Ragoussis (2014). "Integrated analysis of microRNA and mRNA expression and association with HIF binding reveals the complexity of microRNA expression regulation under hypoxia." Mol Cancer **13**: 28.

Carthew, R. W. and E. J. Sontheimer (2009). "Origins and Mechanisms of miRNAs and siRNAs." Cell **136**(4): 642-655.

Ceppi, P., S. Novello, A. Cambieri, M. Longo, V. Monica, M. Lo Iacono, M. Giaj-Levra, S. Saviozzi, M. Volante, M. Papotti and G. Scagliotti (2009). "Polymerase eta mRNA expression predicts survival of non-small cell lung cancer patients treated with platinum-based chemotherapy." Clin Cancer Res **15**(3): 1039-1045.

Cerami, E., J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, A. Jacobsen, C. J. Byrne, M. L. Heuer, E. Larsson, Y. Antipin, B. Reva, A. P. Goldberg, C. Sander and N. Schultz (2012). "The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data." Cancer Discov **2**(5): 401-404.

Chen, C., Y. Zhang, M. M. Loomis, M. P. Upton, P. Lohavanichbutr, J. R. Houck, D. R. Doody, E. Mendez, N. Futran, S. M. Schwartz and P. Wang (2015). "Genome-Wide Loss of Heterozygosity and DNA Copy Number Aberration in HPV-Negative Oral Squamous Cell Carcinoma and Their Associations with Disease-Specific Survival." PLoS One **10**(8): e0135074.

Chen, G., J. Hu, Z. Huang, L. Yang and M. Chen (2016). "MicroRNA-1976 functions as a tumor suppressor and serves as a prognostic indicator in non-small cell lung cancer by directly targeting PLCE1." Biochem Biophys Res Commun **473**(4): 1144-1151.

Cheng, C. F., J. Fan, M. Fedesco, S. Guan, Y. Li, B. Bandyopadhyay, A. M. Bright, D. Yerushalmi, M. Liang, M. Chen, Y. P. Han, D. T. Woodley and W. Li (2008). "Transforming growth factor alpha (TGFalpha)-stimulated secretion of HSP90alpha: using the receptor LRP-1/CD91 to promote human skin cell migration against a TGFbeta-rich environment during wound healing." Mol Cell Biol **28**(10): 3344-3358.

Cho, S., I. Jang, Y. Jun, S. Yoon, M. Ko, Y. Kwon, I. Choi, H. Chang, D. Ryu, B. Lee, V. N. Kim, W. Kim and S. Lee (2013). "MiRGator v3.0: a microRNA portal for deep sequencing, expression profiling and mRNA targeting." Nucleic Acids Res **41**(Database issue): D252-257.

Chou, C. H., N. W. Chang, S. Shrestha, S. D. Hsu, Y. L. Lin, W. H. Lee, C. D. Yang, H. C. Hong, T. Y. Wei, S. J. Tu, T. R. Tsai, S. Y. Ho, T. Y. Jian, H. Y. Wu, P. R. Chen, N. C. Lin, H. T. Huang, T. L. Yang, C. Y. Pai, C. S. Tai, W. L. Chen, C. Y. Huang, C. C. Liu, S. L. Weng, K. W. Liao, W. L. Hsu and H. D. Huang (2016). "miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database." Nucleic Acids Res **44**(D1): D239-247.

Cohen, G. M., X. M. Sun, H. Fearnhead, M. MacFarlane, D. G. Brown, R. T. Snowden and D. Dinsdale (1994). "Formation of large molecular weight fragments of DNA is a key committed step of apoptosis in thymocytes." J Immunol **153**(2): 507-516.

Cordes, K. R., N. T. Sheehy, M. P. White, E. C. Berry, S. U. Morton, A. N. Muth, T. H. Lee, J. M. Miano, K. N. Ivey and D. Srivastava (2009). "miR-145 and miR-143 regulate smooth muscle cell fate and plasticity." Nature **460**(7256): 705-710.

Corsten, M. F., R. Miranda, R. Kasmieh, A. M. Krichevsky, R. Weissleder and K. Shah (2007). "MicroRNA-21 knockdown disrupts glioma growth in vivo and displays synergistic cytotoxicity with neural precursor cell delivered S-TRAIL in human gliomas." Cancer Res **67**(19): 8994-9000.

Creighton, C. J., A. K. Nagaraja, S. M. Hanash, M. M. Matzuk and P. H. Gunaratne (2008). "A bioinformatics tool for linking gene expression profiling results with public databases of microRNA target predictions." RNA **14**(11): 2290-2296.

Crnkovic-Mertens, I., N. Wagener, J. Semzow, E. F. Grone, A. Haferkamp, M. Hohenfellner, K. Butz and F. Hoppe-Seyler (2007). "Targeted inhibition of Livin resensitizes renal cancer cells towards apoptosis." Cell Mol Life Sci **64**(9): 1137-1144.

D'Antonio, M., P. D'Onorio De Meo, M. Pallocca, E. Picardi, A. M. D'Erchia, R. A. Calogero, T. Castrignano and G. Pesole (2015). "RAP: RNA-Seq Analysis Pipeline, a new cloud-based NGS web application." BMC Genomics **16**: S3.

Dambal, S., M. Shah, B. Mihelich and L. Nonn (2015). "The microRNA-183 cluster: the family that plays together stays together." Nucleic Acids Res **43**(15): 7173-7188.

Das, A. V. and R. M. Pillai (2015). "Implications of miR cluster 143/145 as universal anti-oncomiRs and their dysregulation during tumorigenesis." Cancer Cell Int **15**: 92.

Davis, M. P., S. van Dongen, C. Abreu-Goodger, N. Bartonicek and A. J. Enright (2013). "Kraken: a set of tools for quality control and analysis of high-throughput sequence data." Methods **63**(1): 41-49.

Denis, N., A. Kitzis, J. Kruh, F. Dautry and D. Corcos (1991). "Stimulation of methotrexate resistance and dihydrofolate reductase gene amplification by c-myc." Oncogene **6**(8): 1453-1457.

Denli, A. M., B. B. Tops, R. H. Plasterk, R. F. Ketting and G. J. Hannon (2004). "Processing of primary microRNAs by the Microprocessor complex." Nature **432**(7014): 231-235.

Dews, M., A. Homayouni, D. Yu, D. Murphy, C. Sevignani, E. Wentzel, E. E. Furth, W. M. Lee, G. H. Enders, J. T. Mendell and A. Thomas-Tikhonenko (2006). "Augmentation of tumor angiogenesis by a Myc-activated microRNA cluster." Nat Genet **38**(9): 1060-1065.

Di Lisio, L., G. Gomez-Lopez, M. Sanchez-Beato, C. Gomez-Abad, M. E. Rodriguez, R. Villuendas, B. I. Ferreira, A. Carro, D. Rico, M. Mollejo, M. A. Martinez, J. Menarguez, A. Diaz-Alderete, J. Gil, J. C. Cigudosa, D. G. Pisano, M. A. Piris and N. Martinez (2010). "Mantle cell lymphoma: transcriptional regulation by microRNAs." Leukemia **24**(7): 1335-1342.

Di Lisio, L., N. Martinez, S. Montes-Moreno, M. Piris-Villaespesa, M. Sanchez-Beato and M. A. Piris (2012). "The role of miRNAs in the pathogenesis and diagnosis of B-cell lymphomas." Blood **120**(9): 1782-1790.

Diez-Villanueva, A., I. Mallona and M. A. Peinado (2015). "Wanderer, an interactive viewer to explore DNA methylation and gene expression data in human cancer." Epigenetics Chromatin **8**: 22.

Dipple, A. (1995). "DNA adducts of chemical carcinogens." Carcinogenesis **16**(3): 437-441.

Doherty, J. A., M. A. Rossing, K. L. Cushing-Haugen, C. Chen, D. J. Van Den Berg, A. H. Wu, M. C. Pike, R. B. Ness, K. Moysich, G. Chenevix-Trench, J. Beesley, P. M. Webb, J. Chang-Claude, S. Wang-Gohrke, M. T. Goodman, G. Lurie, P. J. Thompson, M. E. Carney, E. Hogdall, S. K. Kjaer, C. Hogdall, E. L. Goode, J. M. Cunningham, B. L. Fridley, R. A. Vierkant, A. Berchuck, P. G. Moorman, J. M. Schildkraut, R. T. Palmieri, D. W. Cramer, K. L. Terry, H. P. Yang, M. Garcia-Closas, S. Chanock, J. Lissowska, H. Song, P. D. Pharoah, M. Shah, B. Perkins, V. McGuire, A. S. Whittemore, R. A. Di Cioccio, A. Gentry-Maharaj, U. Menon, S. A. Gayther, S. J. Ramus, A. Ziogas, W. Brewster, H. Anton-Culver, G. Australian Ovarian Cancer Study Management, S. Australian Cancer, C. L. Pearce and C. Ovarian Cancer Association (2010). "ESR1/SYNE1 polymorphism and invasive epithelial ovarian cancer risk: an Ovarian Cancer Association Consortium study." Cancer Epidemiol Biomarkers Prev **19**(1): 245-250.

Du, P., X. Zhang, C. C. Huang, N. Jafari, W. A. Kibbe, L. Hou and S. M. Lin (2010). "Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis." BMC Bioinformatics **11**: 587.

Duan, J., W. Huang and H. Shi (2016). "Positive expression of KIF20A indicates poor prognosis of glioma patients." Onco Targets Ther **9**: 6741-6749.

Dweep, H., N. Gretz and C. Sticht (2014). "miRWalk database for miRNA-target interactions." Methods Mol Biol **1182**: 289-305.

Dyson, N. (1998). "The regulation of E2F by pRB-family proteins." Genes Dev **12**(15): 2245-2262.

Enright, A. J., B. John, U. Gaul, T. Tuschl, C. Sander and D. S. Marks (2003). "MicroRNA targets in Drosophila." Genome Biol **5**(1): R1.

Eulalio, A., J. Rehwinkel, M. Stricker, E. Huntzinger, S. F. Yang, T. Doerks, S. Dorner, P. Bork, M. Boutros and E. Izaurralde (2007). "Target-specific requirements for enhancers of decapping in miRNA-mediated gene silencing." Genes Dev **21**(20): 2558-2570.

Fabregat, A., K. Sidiropoulos, P. Garapati, M. Gillespie, K. Hausmann, R. Haw, B. Jassal, S. Jupe, F. Korninger, S. McKay, L. Matthews, B. May, M. Milacic, K. Rothfels, V. Shamovsky, M. Webber, J. Weiser, M. Williams, G. Wu, L. Stein, H. Hermjakob and P. D'Eustachio (2016). "The Reactome pathway Knowledgebase." Nucleic Acids Res **44**(D1): D481-487.

Fan, X., X. Zhang, X. Wu, H. Guo, Y. Hu, F. Tang and Y. Huang (2015). "Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos." Genome Biol **16**: 148.

Farazi, T. A., H. M. Horlings, J. J. Ten Hoeve, A. Mihailovic, H. Halfwerk, P. Morozov, M. Brown, M. Hafner, F. Reyal, M. van Kouwenhove, B. Kreike, D. Sie, V. Hovestadt, L. F. Wessels, M. J. van de Vijver and T. Tuschl (2011). "MicroRNA sequence and expression analysis in breast tumors by deep sequencing." Cancer Res **71**(13): 4443-4453.

Farh, K. K., A. Grimson, C. Jan, B. P. Lewis, W. K. Johnston, L. P. Lim, C. B. Burge and D. P. Bartel (2005). "The widespread impact of mammalian MicroRNAs on mRNA repression and evolution." Science **310**(5755): 1817-1821.

Ferlay, J., E. Steliarova-Foucher, J. Lortet-Tieulent, S. Rosso, J. W. Coebergh, H. Comber, D. Forman and F. Bray (2013). "Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012." Eur J Cancer **49**(6): 1374-1403.

Fialkow, L., Y. Wang and G. P. Downey (2007). "Reactive oxygen and nitrogen species as signaling molecules regulating neutrophil function." Free Radic Biol Med **42**(2): 153-164.

Fonseca, N. A., J. Marioni and A. Brazma (2014). "RNA-Seq gene profiling--a systematic empirical comparison." PLoS One **9**(9): e107026.

Fontana, L., M. E. Fiori, S. Albini, L. Cifaldi, S. Giovinazzi, M. Forloni, R. Boldrini, A. Donfrancesco, V. Federici, P. Giacomini, C. Peschle and D. Fruci (2008). "Antagomir-17-5p abolishes the growth of therapy-resistant neuroblastoma through p21 and BIM." PLoS One **3**(5): e2236.

Forbes, S. A., D. Beare, P. Gunasekaran, K. Leung, N. Bindal, H. Boutselakis, M. Ding, S. Bamford, C. Cole, S. Ward, C. Y. Kok, M. Jia, T. De, J. W. Teague, M. R. Stratton, U. McDermott and P. J. Campbell (2015). "COSMIC: exploring the world's knowledge of somatic mutations in human cancer." Nucleic Acids Res **43**(Database issue): D805-811.

Friedlander, M. R., S. D. Mackowiak, N. Li, W. Chen and N. Rajewsky (2012). "miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades." Nucleic Acids Res **40**(1): 37-52.

Friedman, R. C., K. K. Farh, C. B. Burge and D. P. Bartel (2009). "Most mammalian mRNAs are conserved targets of microRNAs." Genome Res **19**(1): 92-105.

Fusco, A. and M. Fedele (2007). "Roles of HMGA proteins in cancer." Nat Rev Cancer **7**(12): 899-910.

Gaidatzis, D., E. van Nimwegen, J. Hausser and M. Zavolan (2007). "Inference of miRNA targets using evolutionary conservation and pathway analysis." BMC Bioinformatics **8**: 69.

Galluzzi, L. and G. Kroemer (2008). "Necroptosis: a specialized pathway of programmed necrosis." Cell **135**(7): 1161-1163.

Gao, J., B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, E. Cerami, C. Sander and N. Schultz (2013). "Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal." Sci Signal **6**(269): pl1.

Ghobrial, I. M., T. E. Witzig and A. A. Adjei (2005). "Targeting apoptosis pathways in cancer therapy." CA Cancer J Clin **55**(3): 178-194.

Giaccia, A. J. and M. B. Kastan (1998). "The complexity of p53 modulation: emerging patterns from divergent signals." Genes Dev **12**(19): 2973-2983.

Golstein, P. and G. Kroemer (2007). "Cell death by necrosis: towards a molecular definition." Trends Biochem Sci **32**(1): 37-43.

Gonzalez-Perez, A. and N. Lopez-Bigas (2011). "Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel." Am J Hum Genet **88**(4): 440-449.

Goos, J. A., V. M. Coupe, B. Diosdado, P. M. Delis-Van Diemen, C. Karga, J. A. Belien, B. Carvalho, M. P. van den Tol, H. M. Verheul, A. A. Geldof, G. A. Meijer, O. S. Hoekstra, R. J. Fijneman and P. E. T. g. DeCoDe (2013). "Aurora kinase A (AURKA) expression in colorectal cancer liver metastasis is associated with poor prognosis." Br J Cancer **109**(9): 2445-2452.

Green, D. R. and G. I. Evan (2002). "A matter of life and death." Cancer Cell **1**(1): 19-30.

Grimson, A., K. K. Farh, W. K. Johnston, P. Garrett-Engele, L. P. Lim and D. P. Bartel (2007). "MicroRNA targeting specificity in mammals: determinants beyond seed pairing." Mol Cell **27**(1): 91-105.

Grishok, A., A. E. Pasquinelli, D. Conte, N. Li, S. Parrish, I. Ha, D. L. Baillie, A. Fire, G. Ruvkun and C. C. Mello (2001). "Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing." Cell **106**(1): 23-34.

Guo, H., N. T. Ingolia, J. S. Weissman and D. P. Bartel (2010). "Mammalian microRNAs predominantly act to decrease target mRNA levels." Nature **466**(7308): 835-840.

Guo, J., Y. Miao, B. Xiao, R. Huan, Z. Jiang, D. Meng and Y. Wang (2009). "Differential expression of microRNA species in human gastric cancer versus non-tumorous tissues." J Gastroenterol Hepatol **24**(4): 652-657.

Hanahan, D. and R. A. Weinberg (2000). "The hallmarks of cancer." Cell **100**(1): 57-70.

Hanahan, D. and R. A. Weinberg (2011). "Hallmarks of cancer: the next generation." Cell **144**(5): 646-674.

Hayashita, Y., H. Osada, Y. Tatematsu, H. Yamada, K. Yanagisawa, S. Tomida, Y. Yatabe, K. Kawahara, Y. Sekido and T. Takahashi (2005). "A polycistronic microRNA cluster, miR-17-92, is overexpressed in human lung cancers and enhances cell proliferation." Cancer Res **65**(21): 9628-9632.

Hayflick, L. (1965). "The Limited in Vitro Lifetime of Human Diploid Cell Strains." Exp Cell Res **37**: 614-636.

Helwak, A., G. Kudla, T. Dudnakova and D. Tollervey (2013). "Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding." Cell **153**(3): 654-665.

Herman, J. G. and S. B. Baylin (2003). "Gene silencing in cancer in association with promoter hypermethylation." N Engl J Med **349**(21): 2042-2054.

Himes, B. E., X. Jiang, P. Wagner, R. Hu, Q. Wang, B. Klanderman, R. M. Whitaker, Q. Duan, J. Lasky-Su, C. Nikolos, W. Jester, M. Johnson, R. A. Panettieri, Jr., K. G. Tantisira, S. T. Weiss and Q. Lu (2014). "RNA-Seq transcriptome profiling identifies CRISPLD2 as a glucocorticoid responsive gene that modulates cytokine function in airway smooth muscle cells." PLoS One **9**(6): e99625.

Hiroki, E., J. Akahira, F. Suzuki, S. Nagase, K. Ito, T. Suzuki, H. Sasano and N. Yaegashi (2010). "Changes in microRNA expression levels correlate with clinicopathological features and prognoses in endometrial serous adenocarcinomas." Cancer Sci **101**(1): 241-249.

Hochberg, Y. and Y. Benjamini (1990). "More powerful procedures for multiple significance testing." Stat Med **9**(7): 811-818.

Hoeijmakers, J. H. (2001). "Genome maintenance mechanisms for preventing cancer." Nature **411**(6835): 366-374.

Hsu, S. D., Y. T. Tseng, S. Shrestha, Y. L. Lin, A. Khaleel, C. H. Chou, C. F. Chu, H. Y. Huang, C. M. Lin, S. Y. Ho, T. Y. Jian, F. M. Lin, T. H. Chang, S. L. Weng, K. W. Liao, I. E. Liao, C. C. Liu and H. D. Huang (2014). "miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions." Nucleic Acids Res **42**(Database issue): D78-85.

Huttenhofer, A. and J. Vogel (2006). "Experimental approaches to identify non-coding RNAs." Nucleic Acids Res **34**(2): 635-646.

Hutvagner, G., J. McLachlan, A. E. Pasquinelli, E. Balint, T. Tuschl and P. D. Zamore (2001). "A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA." Science **293**(5531): 834-838.

Jacobsen, A., J. Silber, G. Harinath, J. T. Huse, N. Schultz and C. Sander (2013). "Analysis of microRNA-target interactions across diverse cancer types." Nat Struct Mol Biol **20**(11): 1325-1332.

Jusufovic, E., M. Rijavec, D. Keser, P. Korosec, E. Sodja, E. Iljazovic, Z. Radojevic and M. Kosnik (2012). "let-7b and miR-126 are down-regulated in tumor tissue and correlate with microvessel density and survival outcomes in non--small--cell lung cancer." PLoS One **7**(9): e45577.

Kanehisa, M., Y. Sato, M. Kawashima, M. Furumichi and M. Tanabe (2016). "KEGG as a reference resource for gene and protein annotation." Nucleic Acids Res **44**(D1): D457-462.

Kaneko, N., K. Miura, Z. Gu, H. Karasawa, S. Ohnuma, H. Sasaki, N. Tsukamoto, S. Yokoyama, A. Yamamura, H. Nagase, C. Shibata, I. Sasaki and A. Horii (2009). "siRNA-mediated knockdown against CDCA1 and KNTC2, both frequently overexpressed in colorectal and gastric cancers, suppresses cell proliferation and induces apoptosis." Biochem Biophys Res Commun **390**(4): 1235-1240.

Kang, W., J. H. Tong, R. W. Lung, Y. Dong, W. Yang, Y. Pan, K. M. Lau, J. Yu, A. S. Cheng and K. F. To (2014). "let-7b/g silencing activates AKT signaling to promote gastric carcinogenesis." J Transl Med **12**: 281.

Karra, H., H. Repo, I. Ahonen, E. Loyttyniemi, R. Pitkanen, M. Lintunen, T. Kuopio, M. Soderstrom and P. Kronqvist (2014). "Cdc20 and securin overexpression predict short-term breast cancer survival." Br J Cancer **110**(12): 2905-2913.

Kawamura, K., R. Bahar, M. Seimiya, M. Chiyo, A. Wada, S. Okada, M. Hatano, T. Tokuhisa, H. Kimura, S. Watanabe, I. Honda, S. Sakiyama, M. Tagawa and O. W. J (2004). "DNA polymerase theta is preferentially expressed in lymphoid tissues and upregulated in human cancers." Int J Cancer **109**(1): 9-16.

Kerr, J. F., A. H. Wyllie and A. R. Currie (1972). "Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics." Br J Cancer **26**(4): 239-257.

Kertesz, M., N. Iovino, U. Unnerstall, U. Gaul and E. Segal (2007). "The role of site accessibility in microRNA target recognition." Nat Genet **39**(10): 1278-1284.

Kielbassa, C., L. Roza and B. Epe (1997). "Wavelength dependence of oxidative DNA damage induced by UV and visible light." Carcinogenesis **18**(4): 811-816.

Kim, B. G., S. Kang, H. H. Han, J. H. Lee, J. E. Kim, S. H. Lee and N. H. Cho (2016). "Transcriptome-wide analysis of compression-induced microRNA expression alteration in breast cancer for mining therapeutic targets." Oncotarget **7**(19): 27468-27478.

Kim, D., G. Pertea, C. Trapnell, H. Pimentel, R. Kelley and S. L. Salzberg (2013). "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions." Genome Biol **14**(4): R36.

Kim, Y. S., M. J. Morgan, S. Choksi and Z. G. Liu (2007). "TNF-induced activation of the Nox1 NADPH oxidase and its role in the induction of necrotic cell death." Mol Cell **26**(5): 675-687.

Kloosterman, W. P., A. K. Lagendijk, R. F. Ketting, J. D. Moulton and R. H. Plasterk (2007). "Targeted inhibition of miRNA maturation with morpholinos reveals a role for miR-375 in pancreatic islet development." PLoS Biol **5**(8): e203.

Kloosterman, W. P. and R. H. Plasterk (2006). "The diverse functions of microRNAs in animal development and disease." Dev Cell **11**(4): 441-450.

Kozomara, A. and S. Griffiths-Jones (2014). "miRBase: annotating high confidence microRNAs using deep sequencing data." Nucleic Acids Res **42**(Database issue): D68-73.

Krek, A., D. Grun, M. N. Poy, R. Wolf, L. Rosenberg, E. J. Epstein, P. MacMenamin, I. da Piedade, K. C. Gunsalus, M. Stoffel and N. Rajewsky (2005). "Combinatorial microRNA target predictions." Nat Genet **37**(5): 495-500.

Krepischi, A. C., M. Maschietto, E. N. Ferreira, A. G. Silva, S. S. Costa, I. W. da Cunha, B. D. Barros, P. E. Grundy, C. Rosenberg and D. M. Carraro (2016). "Genomic imbalances pinpoint potential oncogenes and tumor suppressors in Wilms tumors." Mol Cytogenet **9**: 20.

Krol, J., I. Loedige and W. Filipowicz (2010). "The widespread regulation of microRNA biogenesis, function and decay." Nat Rev Genet **11**(9): 597-610.

Kruger, J. and M. Rehmsmeier (2006). "RNAhybrid: microRNA target prediction easy, fast and flexible." Nucleic Acids Res **34**(Web Server issue): W451-454.

Krutzfeldt, J., N. Rajewsky, R. Braich, K. G. Rajeev, T. Tuschl, M. Manoharan and M. Stoffel (2005). "Silencing of microRNAs in vivo with 'antagomirs'." Nature **438**(7068): 685-689.

Lambie, H., A. Miremadi, S. E. Pinder, J. A. Bell, P. Wencyk, E. C. Paish, R. D. Macmillan and I. O. Ellis (2003). "Prognostic significance of BRCA1 expression in sporadic breast carcinomas." J Pathol **200**(2): 207-213.

Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." Nat Methods **9**(4): 357-359.

Langmead, B., C. Trapnell, M. Pop and S. L. Salzberg (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." Genome Biol **10**(3): R25.

Le Brigand, K., K. Robbe-Sermesant, B. Mari and P. Barbry (2010). "MiRonTop: mining microRNAs targets across large scale gene expression studies." Bioinformatics **26**(24): 3131-3132.

Lee, C. T., T. T. Wu, C. M. Lohse and L. Zhang (2014). "High-mobility group AT-hook 2: an independent marker of poor prognosis in intrahepatic cholangiocarcinoma." Hum Pathol **45**(11): 2334-2340.

Lee, R. C., R. L. Feinbaum and V. Ambros (1993). "The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14." Cell **75**(5): 843-854.

Lee, T. I. and R. A. Young (2013). "Transcriptional regulation and its misregulation in disease." Cell **152**(6): 1237-1251.

Lee, Y., C. Ahn, J. Han, H. Choi, J. Kim, J. Yim, J. Lee, P. Provost, O. Radmark, S. Kim and V. N. Kim (2003). "The nuclear RNase III Drosha initiates microRNA processing." Nature **425**(6956): 415-419.

Lee, Y. C., C. C. Huang, D. Y. Lin, W. C. Chang and K. H. Lee (2015). "Overexpression of centromere protein K (CENPK) in ovarian cancer is correlated with poor patient survival and associated with predictive and prognostic relevance." PeerJ **3**: e1386.

Lesnock, J. L., K. M. Darcy, C. Tian, J. A. Deloia, M. M. Thrall, C. Zahn, D. K. Armstrong, M. J. Birrer and T. C. Krivak (2013). "BRCA1 expression and improved survival in ovarian cancer patients treated with intraperitoneal cisplatin and paclitaxel: a Gynecologic Oncology Group Study." Br J Cancer **108**(6): 1231-1237.

Levine, A. J. (1989). "The p53 tumor suppressor gene and gene product." Princess Takamatsu Symp **20**: 221-230.

Lewis, B. P., C. B. Burge and D. P. Bartel (2005). "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets." Cell **120**(1): 15-20.

Lewis, H., R. Lance, D. Troyer, H. Beydoun, M. Hadley, J. Orians, T. Benzine, K. Madric, O. J. Semmes, R. Drake and A. Esquela-Kerscher (2014). "miR-888 is an expressed prostatic secretions-derived microRNA that promotes prostate cell growth and migration." Cell Cycle **13**(2): 227-239.

Li, Y. and Z. Zhang (2014). "Potential microRNA-mediated oncogenic intercellular communication revealed by pan-cancer analysis." Sci Rep **4**: 7097.

Liang, M. L., T. H. Hsieh, K. H. Ng, Y. N. Tsai, C. F. Tsai, M. E. Chao, D. J. Liu, S. S. Chu, W. Chen, Y. R. Liu, R. S. Liu, S. C. Lin, D. M. Ho, T. T. Wong, M. H. Yang and H. W. Wang (2016). "Downregulation of miR-137 and miR-6500-3p promotes cell proliferation in pediatric high-grade gliomas." Oncotarget **7**(15): 19723-19737.

Liao, Y., G. K. Smyth and W. Shi (2014). "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features." Bioinformatics **30**(7): 923-930.

Lim, L. P., N. C. Lau, P. Garrett-Engele, A. Grimson, J. M. Schelter, J. Castle, D. P. Bartel, P. S. Linsley and J. M. Johnson (2005). "Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs." Nature **433**(7027): 769-773.

Los, M., M. Mozoluk, D. Ferrari, A. Stepczynska, C. Stroh, A. Renz, Z. Herceg, Z. Q. Wang and K. Schulze-Osthoff (2002). "Activation and caspase-mediated inhibition of PARP: a molecular switch between fibroblast necrosis and apoptosis in death receptor signaling." Mol Biol Cell **13**(3): 978-988.

Lu, J., G. Getz, E. A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B. L. Ebert, R. H. Mak, A. A. Ferrando, J. R. Downing, T. Jacks, H. R. Horvitz and T. R. Golub (2005). "MicroRNA expression profiles classify human cancers." Nature **435**(7043): 834-838.

Lv, Y. G., F. Yu, Q. Yao, J. H. Chen and L. Wang (2010). "The role of survivin in diagnosis, prognosis and treatment of breast cancer." J Thorac Dis **2**(2): 100-110.

Lytle, J. R., T. A. Yario and J. A. Steitz (2007). "Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5' UTR as in the 3' UTR." Proc Natl Acad Sci U S A **104**(23): 9667-9672.

Ma, L., G. Z. Li, Z. S. Wu and G. Meng (2014). "Prognostic significance of let-7b expression in breast cancer and correlation to its target gene of BSG expression." Med Oncol **31**(1): 773.

Ma, L., F. Reinhardt, E. Pan, J. Soutschek, B. Bhat, E. G. Marcusson, J. Teruya-Feldstein, G. W. Bell and R. A. Weinberg (2010). "Therapeutic silencing of miR-10b inhibits metastasis in a mouse mammary tumor model." Nat Biotechnol **28**(4): 341-347.

Mahner, S., C. Baasch, J. Schwarz, S. Hein, L. Wolber, F. Janicke and K. Milde-Langosch (2008). "C-Fos expression is a molecular predictor of progression and survival in epithelial ovarian carcinoma." Br J Cancer **99**(8): 1269-1275.

Martin, S. J. and D. R. Green (1995). "Protease activation during apoptosis: death by a thousand cuts?" Cell **82**(3): 349-352.

Martin-Perez, D., P. Vargiu, S. Montes-Moreno, E. A. Leon, S. M. Rodriguez-Pinilla, L. D. Lisio, N. Martinez, R. Rodriguez, M. Mollejo, J. Castellvi, D. G. Pisano, M. Sanchez-Beato and M. A. Piris (2012). "Epstein-Barr virus microRNAs repress BCL6 expression in diffuse large B-cell lymphoma." Leukemia **26**(1): 180-183.

Matamala, N., M. T. Vargas, R. Gonzalez-Campora, J. I. Arias, P. Menendez, E. Andres-Leon, K. Yanowsky, A. Llaneza-Folgueras, R. Minambres, B. Martinez-Delgado and J. Benitez (2016). "MicroRNA deregulation in triple negative breast cancer reveals a role of miR-498 in regulating BRCA1 expression." Oncotarget **7**(15): 20068-20079.

Mei, M., Y. Ren, X. Zhou, X. B. Yuan, L. Han, G. X. Wang, Z. Jia, P. Y. Pu, C. S. Kang and Z. Yao (2010). "Downregulation of miR-21 enhances chemotherapeutic effect of taxol in breast carcinoma cells." Technol Cancer Res Treat **9**(1): 77-86.

Mermel, C. H., S. E. Schumacher, B. Hill, M. L. Meyerson, R. Beroukhim and G. Getz (2011). "GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers." Genome Biol **12**(4): R41.

Mertins, P., D. R. Mani, K. V. Ruggles, M. A. Gillette, K. R. Clauser, P. Wang, X. Wang, J. W. Qiao, S. Cao, F. Petralia, E. Kawaler, F. Mundt, K. Krug, Z. Tu, J. T. Lei, M. L. Gatza, M. Wilkerson, C. M. Perou, V. Yellapantula, K. L. Huang, C. Lin, M. D. McLellan, P. Yan, S. R. Davies, R. R. Townsend, S. J. Skates, J. Wang, B. Zhang, C. R. Kinsinger, M. Mesri, H. Rodriguez,

L. Ding, A. G. Paulovich, D. Fenyo, M. J. Ellis, S. A. Carr and C. Nci (2016). "Proteogenomics connects somatic mutations to signalling in breast cancer." Nature **534**(7605): 55-62.

Meyer, B., S. Loeschke, A. Schultze, T. Weigel, M. Sandkamp, T. Goldmann, E. Vollmer and J. Bullerdiek (2007). "HMGA2 overexpression in non-small cell lung cancer." Mol Carcinog **46**(7): 503-511.

Michaloglou, C., L. C. Vredeveld, M. S. Soengas, C. Denoyelle, T. Kuilman, C. M. van der Horst, D. M. Majoor, J. W. Shay, W. J. Mooi and D. S. Peeper (2005). "BRAFE600-associated senescence-like cell cycle arrest of human naevi." Nature **436**(7051): 720-724.

Mikula, M., J. Gotzmann, A. N. Fischer, M. F. Wolschek, C. Thallinger, R. Schulte-Hermann, H. Beug and W. Mikulits (2003). "The proto-oncoprotein c-Fos negatively regulates hepatocellular tumorigenesis." Oncogene **22**(43): 6725-6738.

Miller, E. C. and J. A. Miller (1981). "Mechanisms of chemical carcinogenesis." Cancer **47**(5 Suppl): 1055-1064.

Min, H. and S. Yoon (2010). "Got target? Computational methods for microRNA target prediction and their extension." Exp Mol Med **42**(4): 233-244.

Morimura, R., S. Komatsu, D. Ichikawa, H. Takeshita, M. Tsujiura, H. Nagata, H. Konishi, A. Shiozaki, H. Ikoma, K. Okamoto, T. Ochiai, H. Taniguchi and E. Otsuji (2011). "Novel diagnostic value of circulating miR-18a in plasma of patients with pancreatic cancer." Br J Cancer **105**(11): 1733-1740.

Morishita, A., M. R. Zaidi, A. Mitoro, D. Sankarasharma, M. Szabolcs, Y. Okada, J. D'Armiento and K. Chada (2013). "HMGA2 is a driver of tumor metastasis." Cancer Res **73**(14): 4289-4299.

Morozova, N., A. Zinovyev, N. Nonne, L. L. Pritchard, A. N. Gorban and A. Harel-Bellan (2012). "Kinetic signatures of microRNA modes of action." RNA **18**(9): 1635-1655.

Nielsen, J. A., P. Lau, D. Maric, J. L. Barker and L. D. Hudson (2009). "Integrating microRNA and mRNA expression profiles of neuronal progenitors to identify regulatory networks underlying the onset of cortical neurogenesis." BMC Neurosci **10**: 98.

Nishida, K., O. Yamaguchi and K. Otsu (2008). "Crosstalk between autophagy and apoptosis in heart disease." Circ Res **103**(4): 343-351.

Nookaew, I., M. Papini, N. Pornputtapong, G. Scalcinati, L. Fagerberg, M. Uhlen and J. Nielsen (2012). "A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae." Nucleic Acids Res **40**(20): 10084-10097.

Ota, A., H. Tagawa, S. Karnan, S. Tsuzuki, A. Karpas, S. Kira, Y. Yoshida and M. Seto (2004). "Identification and characterization of a novel gene, C13orf25, as a target for 13q31-q32 amplification in malignant lymphoma." Cancer Res **64**(9): 3087-3095.

Pasquinelli, A. E., B. J. Reinhart, F. Slack, M. Q. Martindale, M. I. Kuroda, B. Maller, D. C. Hayward, E. E. Ball, B. Degnan, P. Muller, J. Spring, A. Srinivasan, M. Fishman, J. Finnerty, J. Corbo, M. Levine, P. Leahy, E. Davidson and G. Ruvkun (2000). "Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA." Nature **408**(6808): 86-89.

Pau Creixell, J. r. R., Syed Haider, Guanming Wu, Tatsuhiro Shibata, Miguel Vazquez, Ville Mustonen, Abel Gonzalez-Perez, John Pearson, Chris Sander, Benjamin J Raphael, Debora S Marks, B F Francis Ouellette, Alfonso Valencia, Gary D Bader, Paul C Boutros, Joshua M Stuart, Rune Linding, Nuria Lopez-Bigas & Lincoln D Stein (2015). "Pathway and network analysis of cancer genomes." Nat Methods **12**(7): 615-621.

Pei, B., C. Sisu, A. Frankish, C. Howald, L. Habegger, X. J. Mu, R. Harte, S. Balasubramanian, A. Tanzer, M. Diekhans, A. Reymond, T. J. Hubbard, J. Harrow and M. B. Gerstein (2012). "The GENCODE pseudogene resource." Genome Biol **13**(9): R51.

Peng, Y., J. Laser, G. Shi, K. Mittal, J. Melamed, P. Lee and J. J. Wei (2008). "Antiproliferative effects by Let-7 repression of high-mobility group A2 in uterine leiomyoma." Mol Cancer Res **6**(4): 663-673.

Pierce, A. M., R. Schneider-Broussard, I. B. Gimenez-Conti, J. L. Russell, C. J. Conti and D. G. Johnson (1999). "E2F1 has both oncogenic and tumor-suppressive properties in a transgenic model." Mol Cell Biol **19**(9): 6408-6414.

Pillai, R. S., C. G. Artus and W. Filipowicz (2004). "Tethering of human Ago proteins to mRNA mimics the miRNA-mediated repression of protein synthesis." RNA **10**(10): 1518-1525.

Pillaire, M. J., J. Selves, K. Gordien, P. A. Gourraud, C. Gentil, M. Danjoux, C. Do, V. Negre, A. Bieth, R. Guimbaud, D. Trouche, P. Pasero, M. Mechali, J. S. Hoffmann and C. Cazaux (2010). "A 'DNA replication' signature of progression and negative outcome in colorectal cancer." Oncogene **29**(6): 876-887.

Place, R. F., L. C. Li, D. Pookot, E. J. Noonan and R. Dahiya (2008). "MicroRNA-373 induces expression of genes with complementary promoter sequences." Proc Natl Acad Sci U S A **105**(5): 1608-1613.

Poliseno, L., L. Salmena, J. Zhang, B. Carver, W. J. Haveman and P. P. Pandolfi (2010). "A coding-independent function of gene and pseudogene mRNAs regulates tumour biology." Nature **465**(7301): 1033-1038.

Rebbaa, A., X. Zheng, P. M. Chou and B. L. Mirkin (2003). "Caspase inhibition switches doxorubicin-induced apoptosis to senescence." Oncogene **22**(18): 2805-2811.

Reid, G., M. E. Pel, M. B. Kirschner, Y. Y. Cheng, N. Mugridge, J. Weiss, M. Williams, C. Wright, J. J. Edelman, M. P. Vallely, B. C. McCaughan, S. Klebe, H. Brahmbhatt, J. A. MacDiarmid and N. van Zandwijk (2013). "Restoring expression of miR-16: a novel approach to therapy for malignant pleural mesothelioma." Ann Oncol **24**(12): 3128-3135.

Rio-Machin, A., B. I. Ferreira, T. Henry, G. Gomez-Lopez, X. Agirre, S. Alvarez, S. Rodriguez-Perales, F. Prosper, M. J. Calasanz, J. Martinez, R. Fonseca and J. C. Cigudosa (2013). "Downregulation of specific miRNAs in hyperdiploid multiple myeloma mimics the oncogenic effect of IgH translocations occurring in the non-hyperdiploid subtype." Leukemia **27**(4): 925-931.

Ritchie, W., S. Flamant and J. E. Rasko (2009). "Predicting microRNA targets and functions: traps for the unwary." Nat Methods **6**(6): 397-398.

Robinson, M. D., D. J. McCarthy and G. K. Smyth (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." Bioinformatics **26**(1): 139-140.

Robinson, M. D. and A. Oshlack (2010). "A scaling normalization method for differential expression analysis of RNA-seq data." Genome Biol **11**(3): R25.

Robinson, M. D. and G. K. Smyth (2008). "Small-sample estimation of negative binomial dispersion, with applications to SAGE data." Biostatistics **9**(2): 321-332.

Rodriguez-Rodero, S., A. F. Fernandez, J. L. Fernandez-Morera, P. Castro-Santos, G. F. Bayon, C. Ferrero, R. G. Urdinguio, R. Gonzalez-Marquez, C. Suarez, I. Fernandez-Vega, M. F. Fresno Forcelledo, P. Martinez-Camblor, V. Mancikova, E. Castelblanco, M. Perez, P. I. Marron, M. Mendiola, D. Hardisson, P. Santisteban, G. Riesco-Eizaguirre, X. Matias-Guiu, A. Carnero, M. Robledo, E. Delgado-Alvarez, E. Menendez-Torre and M. F. Fraga (2013). "DNA methylation

signatures identify biologically distinct thyroid cancer subtypes." J Clin Endocrinol Metab **98**(7): 2811-2821.

Roos, W. P. and B. Kaina (2013). "DNA damage-induced cell death: from specific DNA lesions to the DNA damage response and apoptosis." Cancer Lett **332**(2): 237-248.

Rupaimoole, R. and F. J. Slack (2017). "MicroRNA therapeutics: towards a new era for the management of cancer and other diseases." Nat Rev Drug Discov.

Sage, J. and A. Ventura (2011). "miR than meets the eye." Genes Dev **25**(16): 1663-1667.

Sandberg, R., J. R. Neilson, A. Sarma, P. A. Sharp and C. B. Burge (2008). "Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites." Science **320**(5883): 1643-1647.

Sasco, A. J., M. B. Secretan and K. Straif (2004). "Tobacco smoking and cancer: a brief review of recent epidemiological evidence." Lung Cancer **45 Suppl 2**: S3-9.

Saumet, A. and C. H. Lecellier (2006). "Anti-viral RNA silencing: do we look like plants?" Retrovirology **3**: 3.

Selbach, M., B. Schwanhausser, N. Thierfelder, Z. Fang, R. Khanin and N. Rajewsky (2008). "Widespread changes in protein synthesis induced by microRNAs." Nature **455**(7209): 58-63.

Serao, N. V., K. R. Delfino, B. R. Southey, J. E. Beever and S. L. Rodriguez-Zas (2011). "Cell cycle and aging, morphogenesis, and response to stimuli genes are individualized biomarkers of glioblastoma progression and survival." BMC Med Genomics **4**: 49.

Shi, M., D. Liu, H. Duan, B. Shen and N. Guo (2010). "Metastasis-related miRNAs, active players in breast cancer invasion, and metastasis." Cancer Metastasis Rev **29**(4): 785-799.

Siepel, A., G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller and D. Haussler (2005). "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes." Genome Res **15**(8): 1034-1050.

Siggelkow, W., D. Boehm, S. Gebhard, M. Battista, I. Sicking, A. Lebrecht, C. Solbach, B. Hellwig, J. Rahnenfuhrer, H. Koelbl, M. Gehrmann, R. Marchan, C. Cadenas, J. G. Hengstler and M. Schmidt (2012). "Expression of aurora kinase A is associated with metastasis-free survival in node-negative breast cancer patients." BMC Cancer **12**: 562.

Singh, S. and A. Suri (2014). "Targeting the testis-specific heat-shock protein 70-2 (HSP70-2) reduces cellular growth, migration, and invasion in renal cell carcinoma cells." Tumour Biol **35**(12): 12695-12706.

Slaby, O., M. Redova, A. Poprach, J. Nekvindova, R. Iliev, L. Radova, R. Lakomy, M. Svoboda and R. Vyzula (2012). "Identification of MicroRNAs associated with early relapse after nephrectomy in renal cell carcinoma patients." Genes Chromosomes Cancer **51**(7): 707-716.

Sun, C., M. Sang, S. Li, X. Sun, C. Yang, Y. Xi, L. Wang, F. Zhang, Y. Bi, Y. Fu and D. Li (2015). "Hsa-miR-139-5p inhibits proliferation and causes apoptosis associated with down-regulation of c-Met." Oncotarget **6**(37): 39756-39792.

Sylvestre, Y., V. De Guire, E. Querido, U. K. Mukhopadhyay, V. Bourdeau, F. Major, G. Ferbeyre and P. Chartrand (2007). "An E2F/miR-20a autoregulatory feedback loop." J Biol Chem **282**(4): 2135-2143.

Takimoto, M., G. Wei, H. Dosaka-Akita, P. Mao, S. Kondo, N. Sakuragi, I. Chiba, T. Miura, N. Itoh, T. Sasao, R. C. Koya, T. Tsukamoto, S. Fujimoto, H. Katoh and N. Kuzumaki (2002).

"Frequent expression of new cancer/testis gene D40/AF15q14 in lung cancers of smokers." Br J Cancer **86**(11): 1757-1762.

Tanic, M., E. Andres, S. M. Rodriguez-Pinilla, I. Marquez-Rodas, M. Cebollero-Presmanes, V. Fernandez, A. Osorio, J. Benitez and B. Martinez-Delgado (2013). "MicroRNA-based molecular classification of non-BRCA1/2 hereditary breast tumours." Br J Cancer **109**(10): 2724-2734.

Tarazona, S., F. Garcia-Alcalde, J. Dopazo, A. Ferrer and A. Conesa (2011). "Differential expression in RNA-seq: a matter of depth." Genome Res **21**(12): 2213-2223.

Taubert, H., M. Kappler, M. Bache, F. Bartel, T. Kohler, C. Lautenschlager, K. Blumke, P. Wurl, H. Schmidt, A. Meye and S. Hauptmann (2005). "Elevated expression of survivin-splice variants predicts a poor outcome for soft-tissue sarcomas patients." Oncogene **24**(33): 5258-5261.

Teng, C. S. (2000). "Protooncogenes as mediators of apoptosis." Int Rev Cytol **197**: 137-202.

Thadani, R. and M. T. Tammi (2006). "MicroTar: predicting microRNA targets from RNA duplexes." BMC Bioinformatics **7 Suppl 5**: S20.

Thiru, P., D. M. Kern, K. L. McKinley, J. K. Monda, F. Rago, K. C. Su, T. Tsinman, D. Yarar, G. W. Bell and I. M. Cheeseman (2014). "Kinetochore genes are coordinately up-regulated in human tumors as part of a FoxM1-related cell division program." Mol Biol Cell **25**(13): 1983-1994.

Thum, T., P. Galuppo, C. Wolf, J. Fiedler, S. Kneitz, L. W. van Laake, P. A. Doevendans, C. L. Mummery, J. Borlak, A. Haverich, C. Gross, S. Engelhardt, G. Ertl and J. Bauersachs (2007). "MicroRNAs in the human heart: a clue to fetal gene reprogramming in heart failure." Circulation **116**(3): 258-267.

Tian, X. J., F. Liu, X. P. Zhang, J. Li and W. Wang (2012). "A two-step mechanism for cell fate decision by coordination of nuclear and mitochondrial p53 activities." PLoS One **7**(6): e38164.

Ventura, A., A. G. Young, M. M. Winslow, L. Lintault, A. Meissner, S. J. Erkeland, J. Newman, R. T. Bronson, D. Crowley, J. R. Stone, R. Jaenisch, P. A. Sharp and T. Jacks (2008). "Targeted deletion reveals essential and overlapping functions of the miR-17 through 92 family of miRNA clusters." Cell **132**(5): 875-886.

Vergoulis, T., I. S. Vlachos, P. Alexiou, G. Georgakilas, M. Maragkakis, M. Reczko, S. Gerangelos, N. Koziris, T. Dalamagas and A. G. Hatzigeorgiou (2012). "TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support." Nucleic Acids Res **40**(Database issue): D222-229.

Virani, S., J. A. Colacino, J. H. Kim and L. S. Rozek (2012). "Cancer epigenetics: a brief review." ILAR J **53**(3-4): 359-369.

Vishnubalaji, R., R. Hamam, M. H. Abdulla, M. A. Mohammed, M. Kassem, O. Al-Obeed, A. Aldahmash and N. M. Alajez (2015). "Genome-wide mRNA and miRNA expression profiling reveal multiple regulatory networks in colorectal cancer." Cell Death Dis **6**: e1614.

Voulgaridou, G. P., I. Anestopoulos, R. Franco, M. I. Panayiotidis and A. Pappa (2011). "DNA damage induced by endogenous aldehydes: current state of knowledge." Mutat Res **711**(1-2): 13-27.

Vu, T. H., T. Li, D. Nguyen, B. T. Nguyen, X. M. Yao, J. F. Hu and A. R. Hoffman (2000). "Symmetric and asymmetric DNA methylation in the human IGF2-H19 imprinted region." Genomics **64**(2): 132-143.

Wang, D., J. Gu, T. Wang and Z. Ding (2014). "OncomiRDB: a database for the experimentally verified oncogenic and tumor-suppressive microRNAs." Bioinformatics **30**(15): 2237-2238.

Wang, K., H. Y. Lim, S. Shi, J. Lee, S. Deng, T. Xie, Z. Zhu, Y. Wang, D. Pocalyko, W. J. Yang, P. A. Rejto, M. Mao, C. K. Park and J. Xu (2013). "Genomic landscape of copy number aberrations enables the identification of oncogenic drivers in hepatocellular carcinoma." Hepatology **58**(2): 706-717.

Wang, T. S., Q. Q. Ding, R. H. Guo, H. Shen, J. Sun, K. H. Lu, S. H. You, H. M. Ge, Y. Q. Shu and P. Liu (2010). "Expression of livin in gastric cancer and induction of apoptosis in SGC-7901 cells by shRNA-mediated silencing of livin gene." Biomed Pharmacother **64**(5): 333-338.

Wang, Z., L. Xu, Y. Hu, Y. Huang, Y. Zhang, X. Zheng, S. Wang, Y. Wang, Y. Yu, M. Zhang, K. Yuan and W. Min (2016). "miRNA let-7b modulates macrophage polarization and enhances tumor-associated macrophages to promote angiogenesis and mobility in prostate cancer." Sci Rep **6**: 25602.

Weber, M., I. Hellmann, M. B. Stadler, L. Ramos, S. Paabo, M. Rebhan and D. Schubeler (2007). "Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome." Nat Genet **39**(4): 457-466.

Weeraratne, S. D., V. Amani, N. Teider, J. Pierre-Francois, D. Winter, M. J. Kye, S. Sengupta, T. Archer, M. Remke, A. H. Bai, P. Warren, S. M. Pfister, J. A. Steen, S. L. Pomeroy and Y. J. Cho (2012). "Pleiotropic effects of miR-183~96~182 converge to regulate cell survival, proliferation and migration in medulloblastoma." Acta Neuropathol **123**(4): 539-552.

Wei, J. J., X. Wu, Y. Peng, G. Shi, O. Basturk, X. Yang, G. Daniels, I. Osman, J. Ouyang, E. Hernando, A. Pellicer, J. S. Rhim, J. Melamed and P. Lee (2011). "Regulation of HMGA1 expression by microRNA-296 affects prostate cancer growth and invasion." Clin Cancer Res **17**(6): 1297-1305.

Williams, G. H. and K. Stoeber (2012). "The cell cycle and cancer." J Pathol **226**(2): 352-364.

Williamson, V., A. Kim, B. Xie, G. O. McMichael, Y. Gao and V. Vladimirov (2013). "Detecting miRNAs in deep-sequencing data: a software performance comparison and evaluation." Brief Bioinform **14**(1): 36-45.

Willimott, S. and S. D. Wagner (2012). "Stromal cells and CD40 ligand (CD154) alter the miRNome and induce miRNA clusters including, miR-125b/miR-99a/let-7c and miR-17-92 in chronic lymphocytic leukaemia." Leukemia **26**(5): 1113-1116.

Woods, K., J. M. Thomson and S. M. Hammond (2007). "Direct regulation of an oncogenic micro-RNA cluster by E2F transcription factors." J Biol Chem **282**(4): 2130-2134.

Wu, W. J., K. S. Hu, D. S. Wang, Z. L. Zeng, D. S. Zhang, D. L. Chen, L. Bai and R. H. Xu (2013). "CDC20 overexpression predicts a poor prognosis for patients with colorectal cancer." J Transl Med **11**: 142.

Wu, W. Y., X. Y. Xue, Z. J. Chen, S. L. Han, Y. P. Huang, L. F. Zhang, G. B. Zhu and X. Shen (2011). "Potentially predictive microRNAs of gastric cancer with metastasis to lymph node." World J Gastroenterol **17**(31): 3645-3651.

Xiao, F., Z. Zuo, G. Cai, S. Kang, X. Gao and T. Li (2009). "miRecords: an integrated resource for microRNA-target interactions." Nucleic Acids Res **37**(Database issue): D105-110.

Xie, B., Q. Ding, H. Han and D. Wu (2013). "miRCancer: a microRNA-cancer association database constructed by text mining on literature." Bioinformatics **29**(5): 638-644.

Xu, Y. and D. Baltimore (1996). "Dual roles of ATM in the cellular response to radiation and in cell growth control." Genes Dev **10**(19): 2401-2410.

Yamasaki, L. (1998). "Growth regulation by the E2F and DP transcription factor families." Results Probl Cell Differ **22**: 199-227.

Yamazaki, H., T. Mori, M. Yazawa, A. M. Maeshima, F. Matsumoto, S. Yoshimoto, Y. Ota, A. Kaneko, H. Tsuda and Y. Kanai (2013). "Stem cell self-renewal factors Bmi1 and HMGA2 in head and neck squamous cell carcinoma: clues for diagnosis." Lab Invest **93**(12): 1331-1338.

Yan, B. (2011). "Research progress on Livin protein: an inhibitor of apoptosis." Mol Cell Biochem **357**(1-2): 39-45.

Yates, A., W. Akanni, M. R. Amode, D. Barrell, K. Billis, D. Carvalho-Silva, C. Cummins, P. Clapham, S. Fitzgerald, L. Gil, C. G. Giron, L. Gordon, T. Hourlier, S. E. Hunt, S. H. Janacek, N. Johnson, T. Juettemann, S. Keenan, I. Lavidas, F. J. Martin, T. Maurel, W. McLaren, D. N. Murphy, R. Nag, M. Nuhn, A. Parker, M. Patricio, M. Pignatelli, M. Rahtz, H. S. Riat, D. Sheppard, K. Taylor, A. Thormann, A. Vullo, S. P. Wilder, A. Zadissa, E. Birney, J. Harrow, M. Muffato, E. Perry, M. Ruffier, G. Spudich, S. J. Trevanion, F. Cunningham, B. L. Aken, D. R. Zerbino and P. Flicek (2016). "Ensembl 2016." Nucleic Acids Res **44**(D1): D710-716.

Yi, R., Y. Qin, I. G. Macara and B. R. Cullen (2003). "Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs." Genes Dev **17**(24): 3011-3016.

Yonemori, M., N. Seki, H. Yoshino, R. Matsushita, K. Miyamoto, M. Nakagawa and H. Enokida (2016). "Dual tumor-suppressors miR-139-5p and miR-139-3p targeting matrix metalloprotease 11 (MMP11) in bladder cancer." Cancer Sci.

Yoshizawa, K., E. Jelezcova, A. R. Brown, J. F. Foley, A. Nyska, X. Cui, L. J. Hofseth, R. M. Maronpot, S. H. Wilson, A. R. Sepulveda and R. W. Sobol (2009). "Gastrointestinal hyperplasia with altered expression of DNA polymerase beta." PLoS One **4**(8): e6493.

Young, M. D., M. J. Wakefield, G. K. Smyth and A. Oshlack (2010). "Gene ontology analysis for RNA-seq: accounting for selection bias." Genome Biol **11**(2): R14.

Zhang, B., J. Wang, X. Wang, J. Zhu, Q. Liu, Z. Shi, M. C. Chambers, L. J. Zimmerman, K. F. Shaddox, S. Kim, S. R. Davies, S. Wang, P. Wang, C. R. Kinsinger, R. C. Rivers, H. Rodriguez, R. R. Townsend, M. J. Ellis, S. A. Carr, D. L. Tabb, R. J. Coffey, R. J. Slebos, D. C. Liebler and C. Nci (2014). "Proteogenomic characterization of human colon and rectal cancer." Nature **513**(7518): 382-387.

Zhang, J., B. Li, Q. Yang, P. Zhang and H. Wang (2015). "Prognostic value of Aurora kinase A (AURKA) expression among solid tumor patients: a systematic review and meta-analysis." Jpn J Clin Oncol **45**(7): 629-636.

Zhang, S., W. Tang, S. Weng, X. Liu, B. Rao, J. Gu, S. Chen, Q. Wang, X. Shen, R. Xue and L. Dong (2014). "Apollon modulates chemosensitivity in human esophageal squamous cell carcinoma." Oncotarget **5**(16): 7183-7197.

Zhang, Y., X. Wen, X. L. Hu, L. Z. Cheng, J. Y. Yu and Z. B. Wei (2016). "Downregulation of miR-145-5p correlates with poor prognosis in gastric cancer." Eur Rev Med Pharmacol Sci **20**(14): 3026-3030.

Zhou, R., X. Zhou, Z. Yin, J. Guo, T. Hu, S. Jiang, L. Liu, X. Dong, S. Zhang and G. Wu (2015). "Tumor invasion and metastasis regulated by microRNA-184 and microRNA-574-5p in small-cell lung cancer." Oncotarget **6**(42): 44609-44622.

Zhu, J., D. Woods, M. McMahon and J. M. Bishop (1998). "Senescence of human fibroblasts induced by oncogenic Raf." Genes Dev **12**(19): 2997-3007.

ANEXO I

TABLAS Y FIGURAS SUPLEMENTARIAS.

Tabla Suplementaria 1. Genes seleccionados de las rutas relacionadas con el cáncer. Contiene un total de 1264 genes (861 únicos) catalogados según las 7 rutas seleccionadas y obtenidos de las bases de datos de Reactome, KEGG y DDRProt. Debido al tamaño de la tabla, solo aparece disponible en la versión electrónica provista en el CD.

Tabla Suplementaria 2. Valores de enriquecimiento funcional en las siete rutas características del cáncer en cada uno de los 15 tipos tumorales.
Los valores señalados en azul, corresponden a las rutas significativas, p-value <= 0.05. Como control, se escogió de forma aleatoria (*Random)* del total de genes desregulados, un número de genes idéntico al de cada una de las siete rutas. Las abreviaturas de los tumores corresponden a: Tumor Cromófobo de riñón (KICH), Carcinoma de cuello y cabeza (HNSC), Carcinoma Esofágico (ESCA), Carcinoma de riñón de célula papilar (KIRP), Carcinoma hepático (LIHC), Carcinoma de riñón de célula clara (KIRC), Adenocarcinoma de pulmón (LUAD), Carcinoma de Tiroides (THAD), Carcinoma de próstata (PRAD), Carcinoma de vejiga (BLCA), Carcinoma de mama (BRCA), Carcinoma escamoso de pulmón (LUSC), Adenocarcinoma de estómago (STAD), Carcinoma de Útero (UCEC) y Cholangiocarcinoma (CHOL). Debido al tamaño de esta tabla, solo está disponible en la versión electrónica provista en el CD.

Tabla Suplementaria 3. Genes diferencialmente expresados pertenecientes a las rutas relacionadas con el cáncer en cada uno de los 15 tipos de tumores.
Cada tabla nuestra el nombre del gen, la ruta a la cual pertenece, el logaritmo del valor de cambio de expresión en base 2, el log2 de las secuencias por cada millón, el valor de probabilidad y el valor ajustado de probabilidad o falso ratio de descubrimiento (FDR). Las abreviaturas de los tumores corresponden a: Tumor Cromófobo de riñón (KICH), Carcinoma de cuello y cabeza (HNSC), Carcinoma Esofágico (ESCA), Carcinoma de riñón de célula papilar (KIRP), Carcinoma hepático (LIHC), Carcinoma de riñón de célula clara (KIRC), Adenocarcinoma de pulmón (LUAD), Carcinoma de Tiroides (THAD), Carcinoma de próstata (PRAD), Carcinoma de vejiga (BLCA), Carcinoma de mama (BRCA), Carcinoma escamoso de pulmón (LUSC), Adenocarcinoma de estómago (STAD), Carcinoma de Útero (UCEC) y Cholangiocarcinoma (CHOL). Debido al tamaño de esta tabla, solo está disponible en la versión electrónica provista en el CD.

Tabla Suplementaria 4. microARNs diferencialmente expresados en cada uno de los 15 tipos de tumores.
Cada tabla nuestra el nombre del miARN, el logaritmo del valor de cambio de expresión en base 2, el log2 de las secuencias por cada millón, el valor de probabilidad y el valor ajustado de probabilidad o falso ratio de descubrimiento (FDR). Debido al tamaño de esta tabla, solo está disponible en la versión electrónica provista en el CD.

Figura Suplementaria 1. Genes y miARNs con mayor cambio de expresión en carcinoma de vejiga. a| La figura muestra los 10 genes con mayor cambio de expresión, 5 sobre-expresados y 5 reprimidos. b| | La figura muestra los 10 microARNs con mayor cambio de expresión, 5 sobre-expresados y 5 inhibidos.

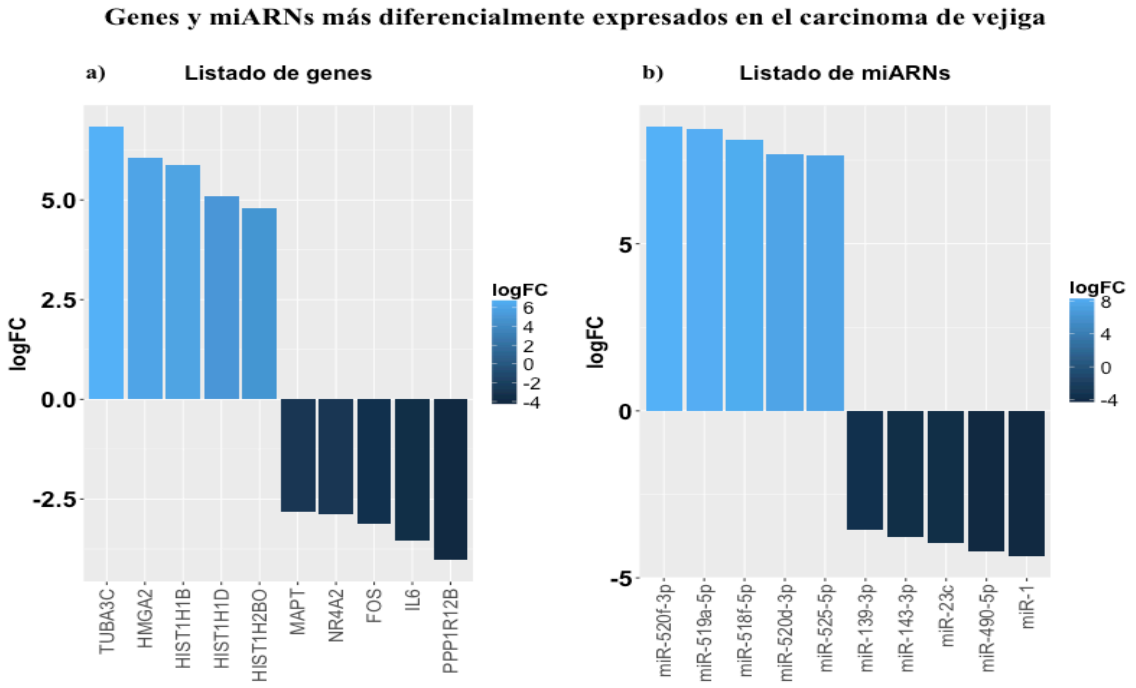


Figura Suplementaria 2. Genes y miARNs con mayor cambio de expresión en carcinoma de mama. a| La figura muestra los 10 genes con mayor cambio de expresión, 5 sobre-expresados y 5 reprimidos. b| | La figura muestra los 10 microARNs con mayor cambio de expresión, 5 sobre-expresados y 5 inhibidos.

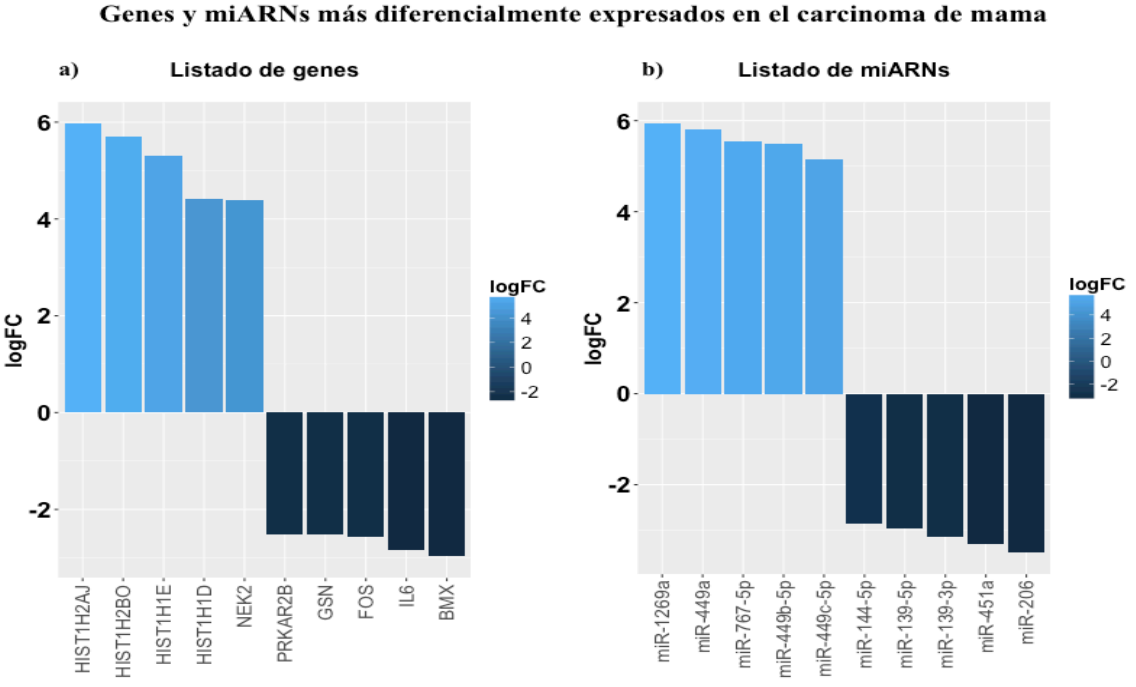


Figura Suplementaria 3. Genes y miARNs con mayor cambio de expresión en cholangiocarcinoma. a| La figura muestra los 10 genes con mayor cambio de expresión, 5 sobre-expresados y 5 reprimidos. b| | La figura muestra los 10 microARNs con mayor cambio de expresión, 5 sobre-expresados y 5 inhibidos.

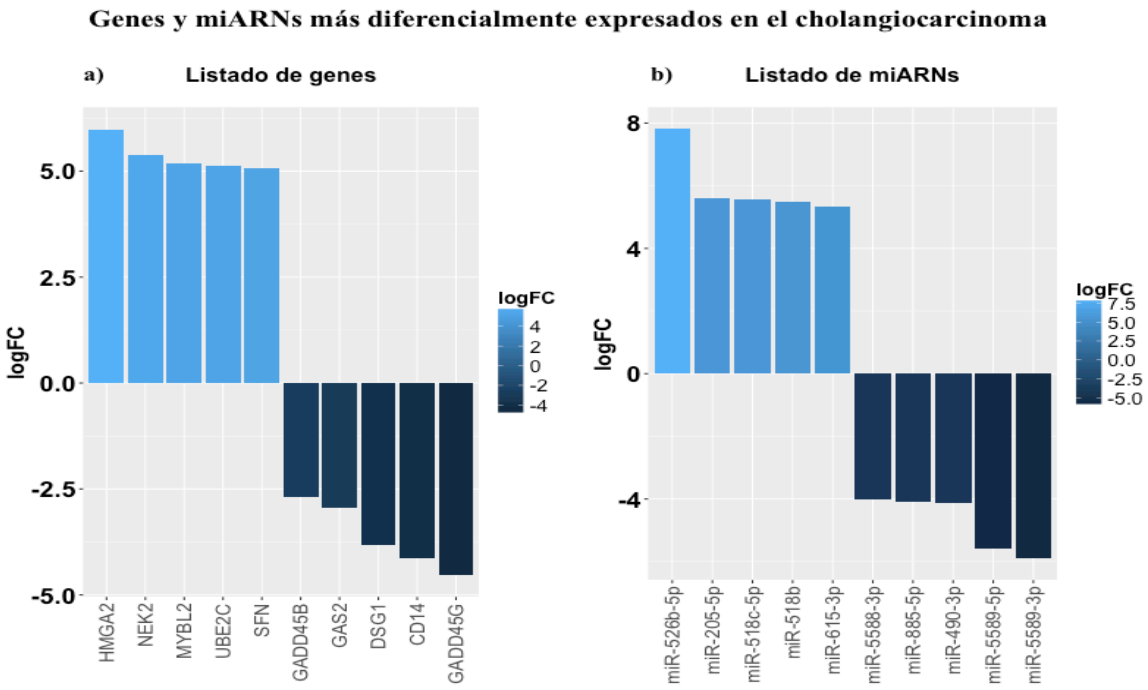


Figura Suplementaria 4. Genes y miARNs con mayor cambio de expresión en carcinoma esofágico. a| La figura muestra los 10 genes con mayor cambio de expresión, 5 sobre-expresados y 5 reprimidos. b| | La figura muestra los 10 microARNs con mayor cambio de expresión, 5 sobre-expresados y 5 inhibidos.

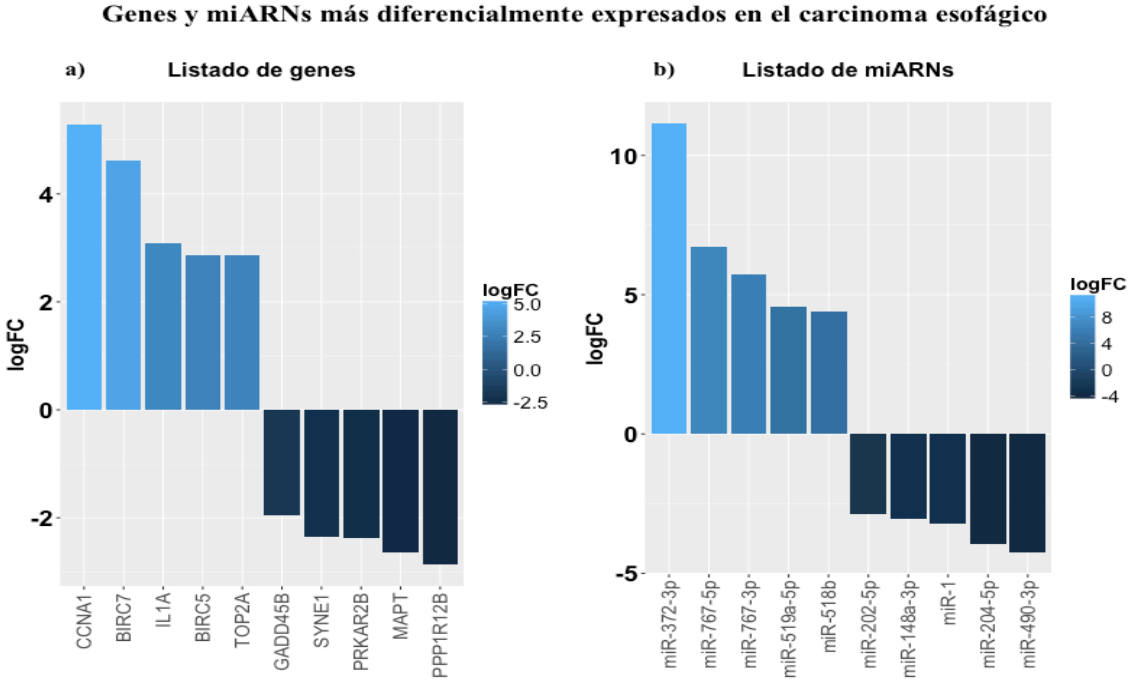


Figura Suplementaria 5. Genes y miARNs con mayor cambio de expresión en tumor de cabeza y cuello. a| La figura muestra los 10 genes con mayor cambio de expresión, 5 sobre-expresados y 5 reprimidos. b| | La figura muestra los 10 microARNs con mayor cambio de expresión, 5 sobre-expresados y 5 inhibidos.

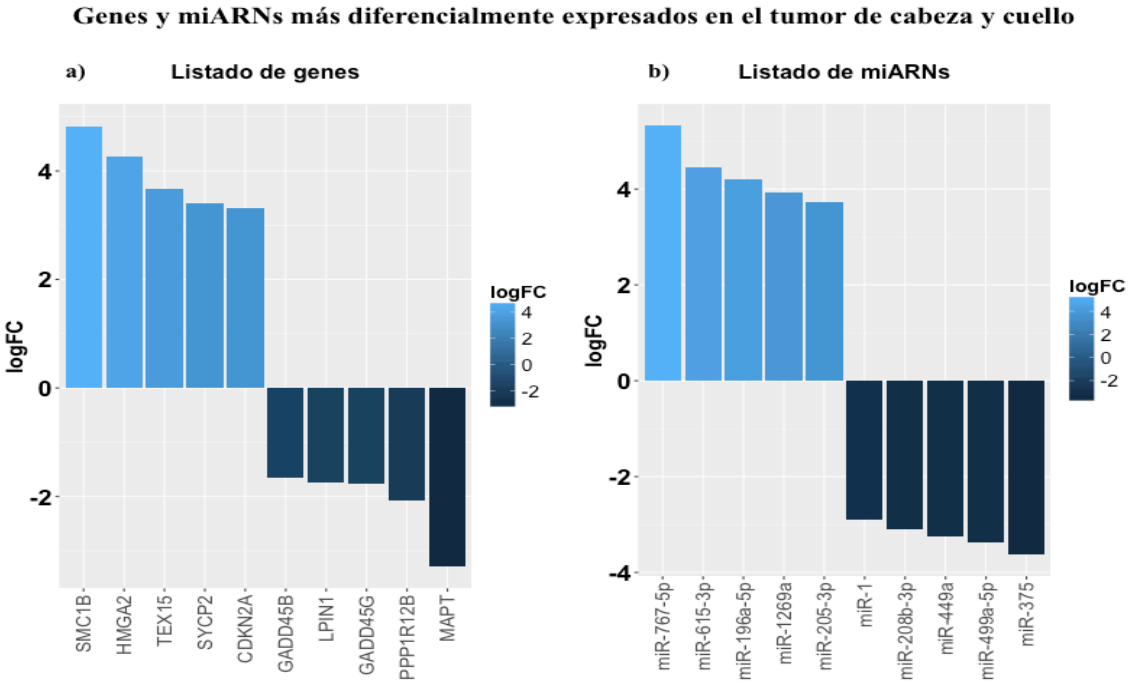


Figura Suplementaria 6. Genes y miARNs con mayor cambio de expresión en tumor cromófobo de riñón. a| La figura muestra los 10 genes con mayor cambio de expresión, 5 sobre-expresados y 5 reprimidos. b| | La figura muestra los 10 microARNs con mayor cambio de expresión, 5 sobre-expresados y 5 inhibidos.

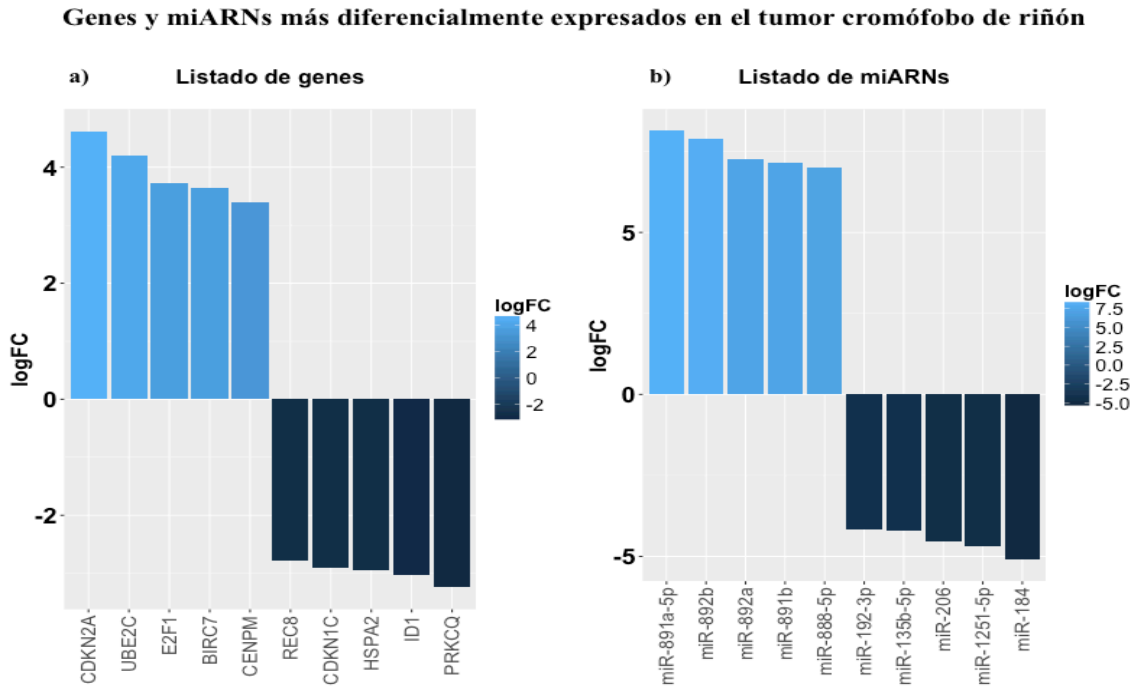


Figura Suplementaria 7. Genes y miARNs con mayor cambio de expresión en tumor de riñón de célula clara. a| La figura muestra los 10 genes con mayor cambio de expresión, 5 sobre-expresados y 5 reprimidos. b| | La figura muestra los 10 microARNs con mayor cambio de expresión, 5 sobre-expresados y 5 inhibidos.

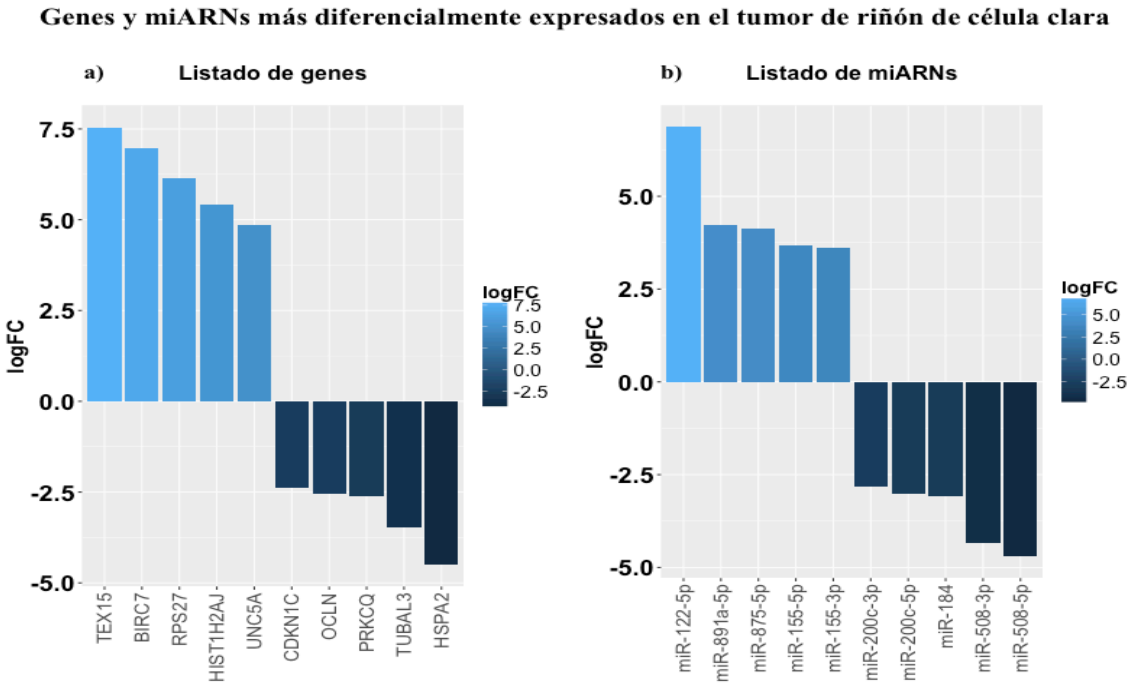


Figura Suplementaria 8. Genes y miARNs con mayor cambio de expresión en tumor de riñón de célula papilar. a| La figura muestra los 10 genes con mayor cambio de expresión, 5 sobre-expresados y 5 reprimidos. b| | La figura muestra los 10 microARNs con mayor cambio de expresión, 5 sobre-expresados y 5 inhibidos.

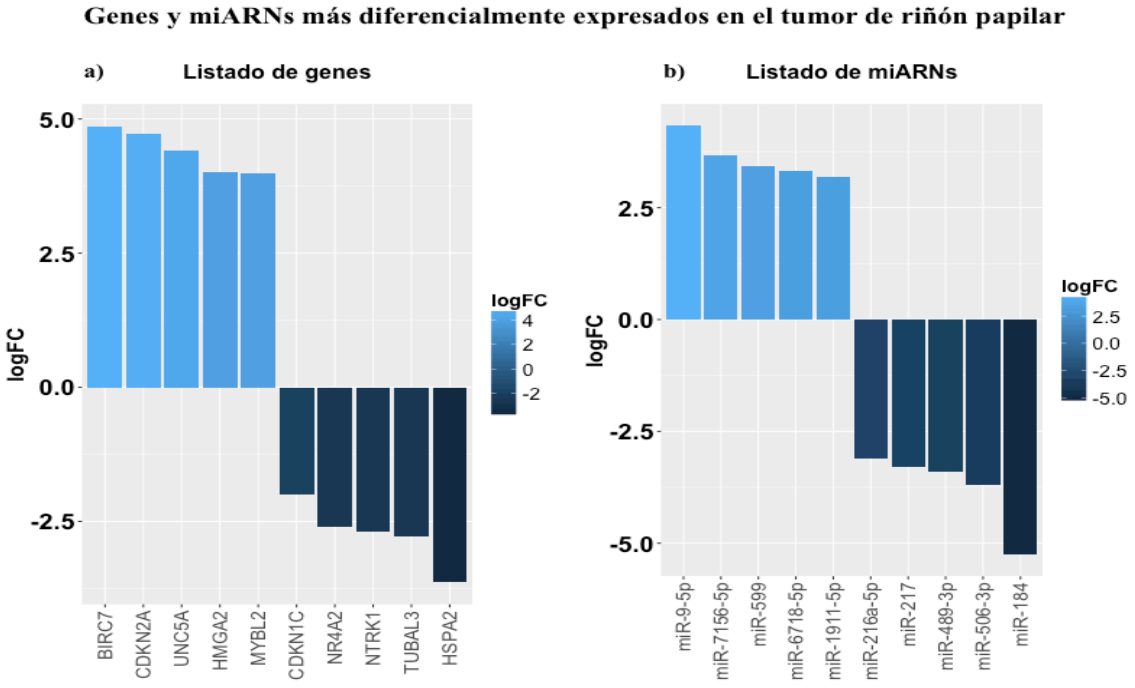


Figura Suplementaria 9. Genes y miARNs con mayor cambio de expresión en carcinoma de hígado. a| La figura muestra los 10 genes con mayor cambio de expresión, 5 sobre-expresados y 5 reprimidos. b| | La figura muestra los 10 microARNs con mayor cambio de expresión, 5 sobre-expresados y 5 inhibidos.

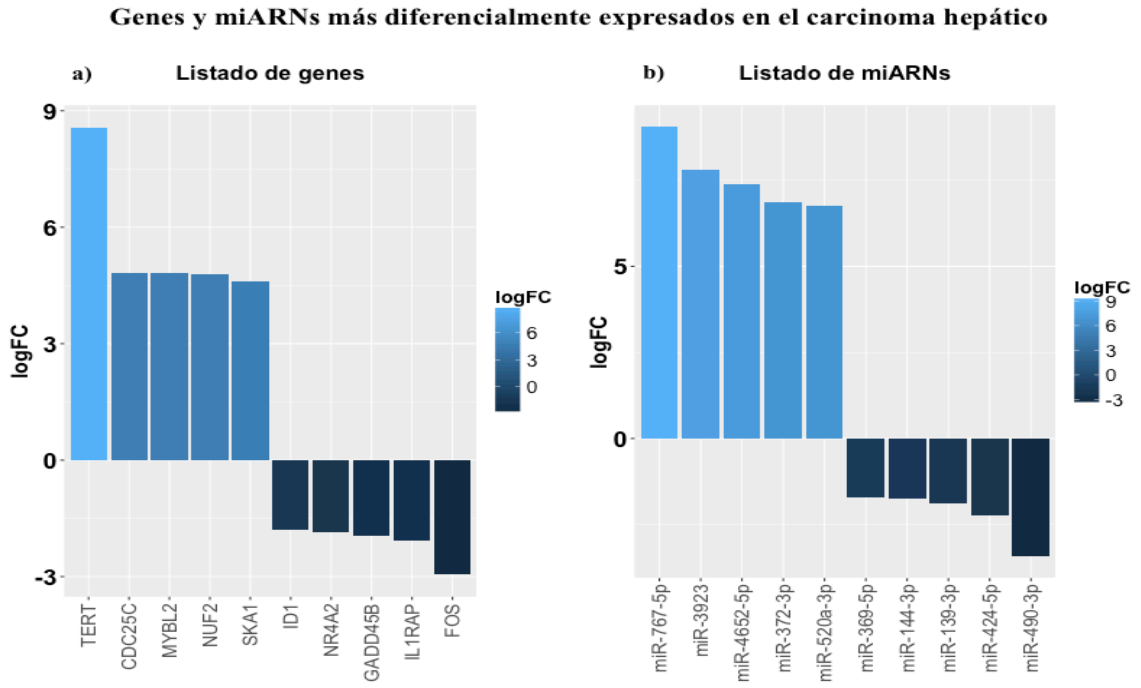


Figura Suplementaria 10. Genes y miARNs con mayor cambio de expresión en adenocarcinoma de pulmón. a| La figura muestra los 10 genes con mayor cambio de expresión, 5 sobre-expresados y 5 reprimidos. b| | La figura muestra los 10 microARNs con mayor cambio de expresión, 5 sobre-expresados y 5 inhibidos.

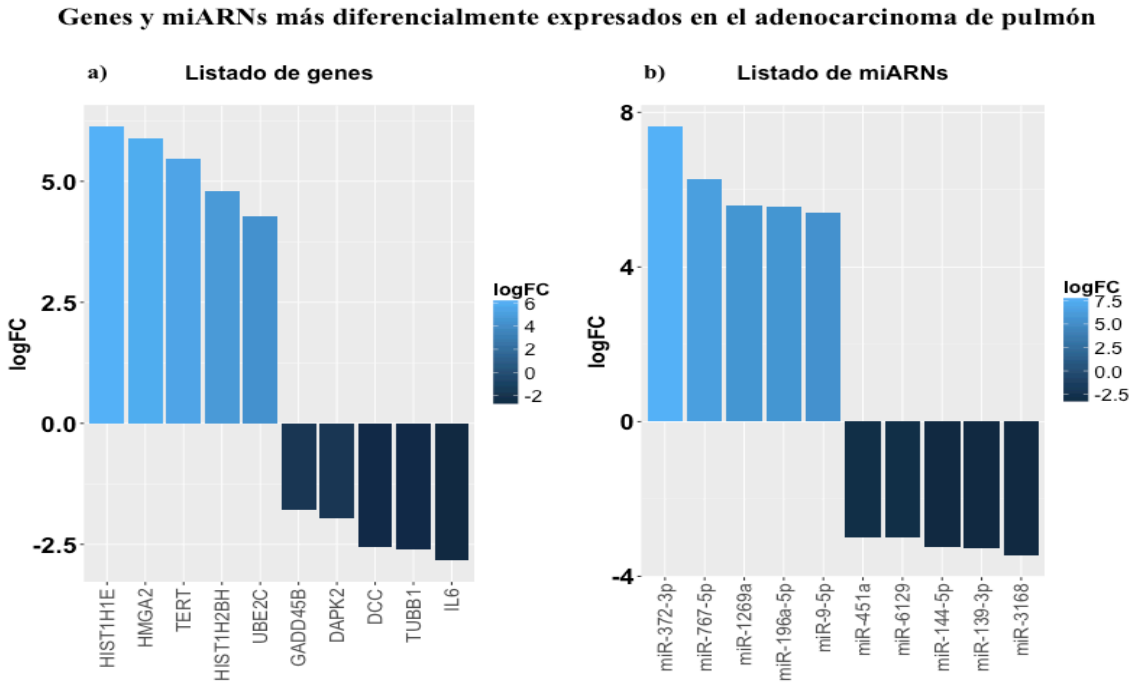


Figura Suplementaria 11. Genes y miARNs con mayor cambio de expresión en carcinoma escamoso de pulmón. a| La figura muestra los 10 genes con mayor cambio de expresión, 5 sobre-expresados y 5 reprimidos. b| | La figura muestra los 10 microARNs con mayor cambio de expresión, 5 sobre-expresados y 5 inhibidos.

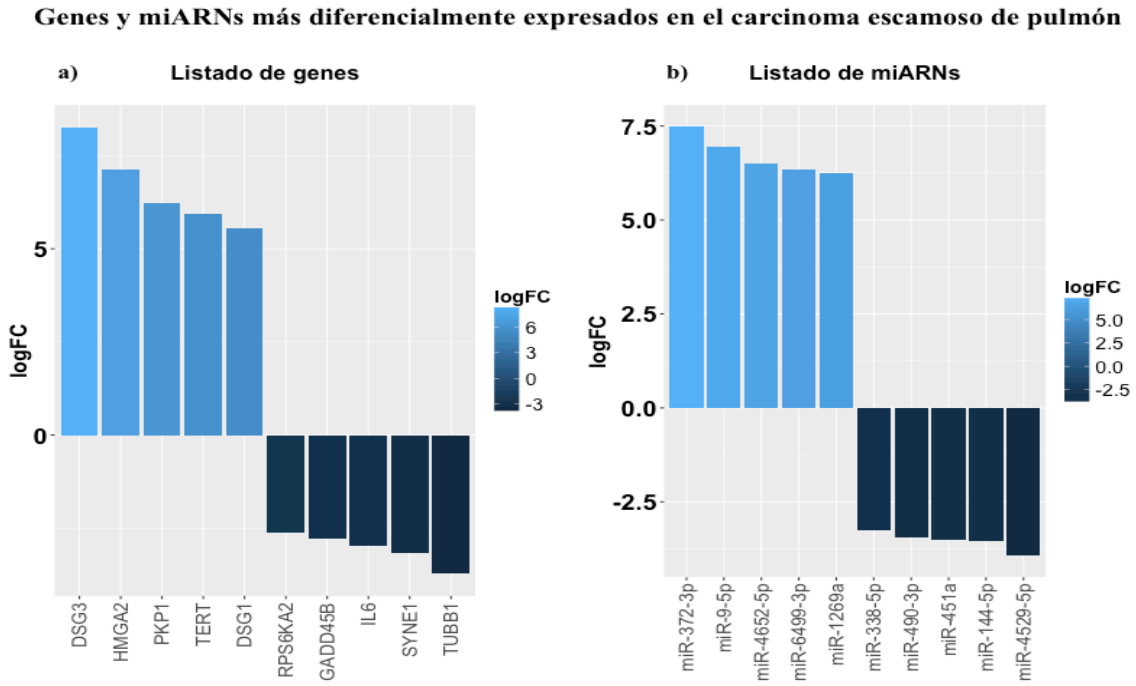


Figura Suplementaria 12. Genes y miARNs con mayor cambio de expresión en adenocarcinoma de próstata. a| La figura muestra los 10 genes con mayor cambio de expresión, 5 sobre-expresados y 5 reprimidos. b| | La figura muestra los 10 microARNs con mayor cambio de expresión, 5 sobre-expresados y 5 inhibidos.

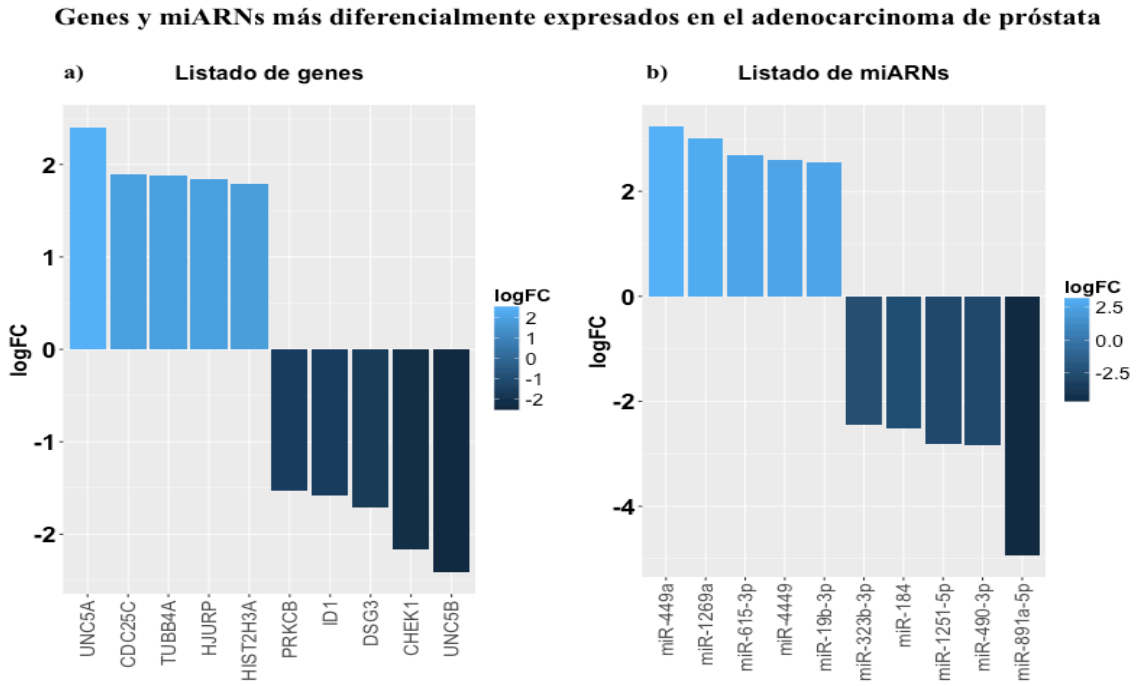


Figura Suplementaria 13. Genes y miARNs con mayor cambio de expresión en adenocarcinoma de estómago. a| La figura muestra los 10 genes con mayor cambio de expresión, 5 sobre-expresados y 5 reprimidos. b| | La figura muestra los 10 microARNs con mayor cambio de expresión, 5 sobre-expresados y 5 inhibidos.

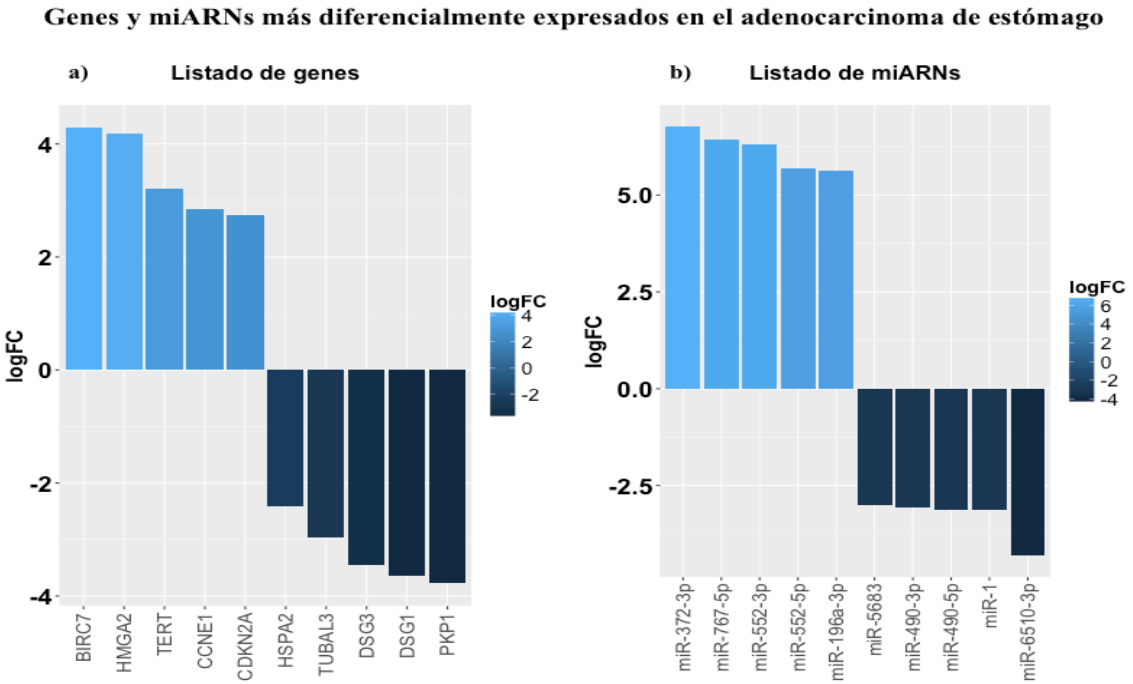


Figura Suplementaria 14. Genes y miARNs con mayor cambio de expresión en carcinoma de tiroides. a| La figura muestra los 10 genes con mayor cambio de expresión, 5 sobre-expresados y 5 reprimidos. b| | La figura muestra los 10 microARNs con mayor cambio de expresión, 5 sobre-expresados y 5 inhibidos.

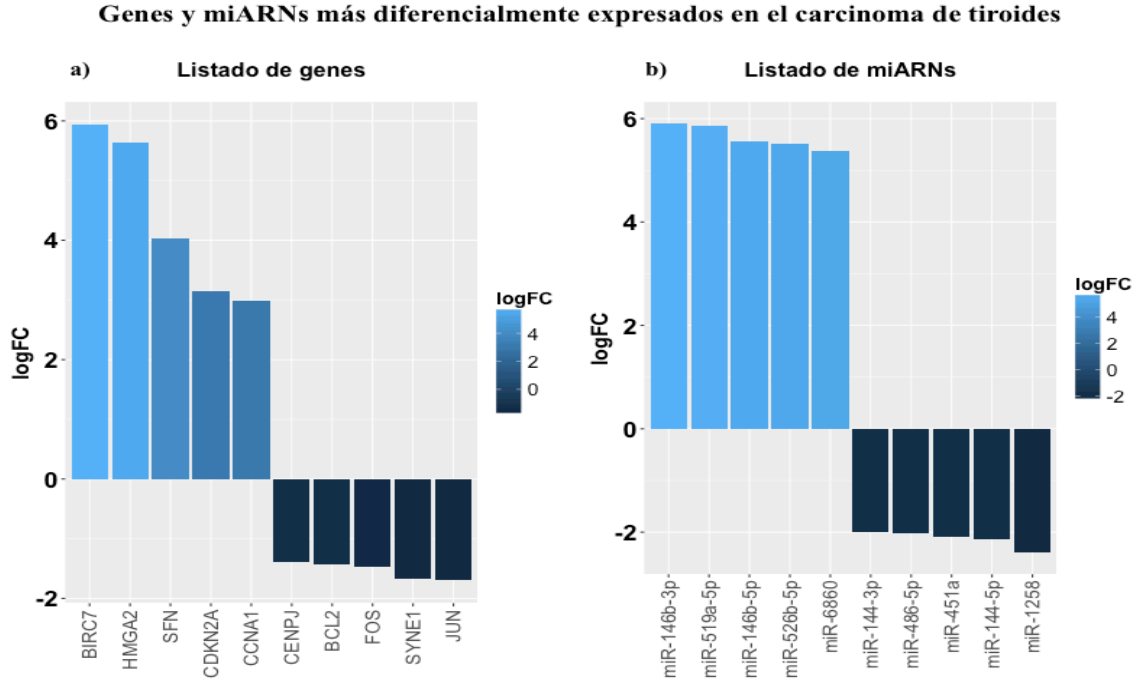


Figura Suplementaria 15. Genes y miARNs con mayor cambio de expresión en carcinoma de útero. a| La figura muestra los 10 genes con mayor cambio de expresión, 5 sobre-expresados y 5 reprimidos. b| | La figura muestra los 10 microARNs con mayor cambio de expresión, 5 sobre-expresados y 5 inhibidos.

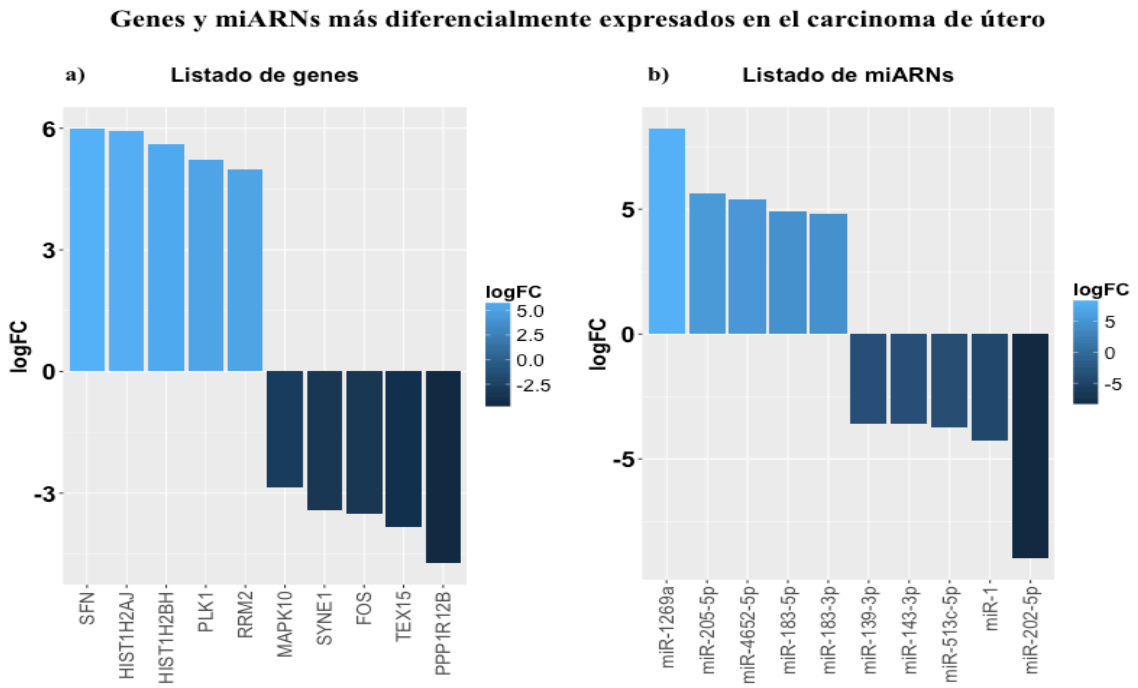


Tabla Suplementaria 5. Genes diferencialmente expresados en la mayoría de los tumores estudiados.

Incluye el nombre del gen y la ruta a la cual pertenece, además el número de tumores donde aparece diferencialmente expresado (Columna nDE), el número de tumores donde aparece sobre-expresado (nDE Up), y donde aparece reprimido (columna nDE Down) y en las dos columnas siguientes los tumores abreviados: Tumor Cromófobo de riñón (KICH), Carcinoma de cuello y cabeza (HNSC), Carcinoma Esofágico (ESCA), Carcinoma de riñón de célula papilar (KIRP), Carcinoma hepático (LIHC), Carcinoma de riñón de célula clara (KIRC), Adenocarcinoma de pulmón (LUAD), Carcinoma de Tiroides (THAD), Carcinoma de próstata (PRAD), Carcinoma de vejiga (BLCA), Carcinoma de mama (BRCA), Carcinoma escamoso de pulmón (LUSC), Carcinoma de estómago (STAD), Carcinoma de Útero (UCEC) y Cholangiocarcinoma (CHOL). Debido al tamaño de esta tabla, solo está disponible en la versión electrónica provista en el CD.

Tabla Suplementaria 6. microARNs diferencialmente expresados en la mayoría de los tumores estudiados.

Esta tabla incluye el nombre del microARN, el número de tumores donde aparece diferencialmente expresado (Columna nDE), el número de tumores en el que aparece sobre-expresado (nDE Up), y donde aparece reprimido (columna nDE Down) y en las dos columnas siguientes, se detallan los tumores abreviados : Tumor Cromófobo de riñón (KICH), Carcinoma de cuello y cabeza (HNSC), Carcinoma Esofágico (ESCA), Carcinoma de riñón de célula papilar (KIRP), Carcinoma hepático (LIHC), Carcinoma de riñón de célula clara (KIRC), Adenocarcinoma de pulmón (LUAD), Carcinoma de Tiroides (THAD), Carcinoma de próstata (PRAD), Carcinoma de vejiga (BLCA), Carcinoma de mama (BRCA), Carcinoma escamoso de pulmón (LUSC), Adenocarcinoma de

estómago (STAD), Carcinoma de Útero (UCEC) y Cholangiocarcinoma (CHOL). Debido al tamaño de esta tabla, solo está disponible en la versión electrónica provista en el CD.

Tabla Suplementaria 7. Relaciones miARN-ARNm conservadas en la mayoría de los tumores.

Incluye el nombre del gen y miARN, el número total de tumores en los que se conserva el par, la expresión del gen y del miARN (+ indica sobreexpresión y - represión) y los nombres de los tumores. La última columna incluye aquellas interacciones publicadas por otros autores, NE significa No Encontrado. Las abreviaturas de los tumores corresponden a: Tumor Cromófobo de riñón (KICH), Carcinoma de cuello y cabeza (HNSC), Carcinoma Esofágico (ESCA), Carcinoma de riñón de célula papilar (KIRP), Carcinoma hepático (LIHC), Carcinoma de riñón de célula clara (KIRC), Adenocarcinoma de pulmón (LUAD), Carcinoma de Tiroides (THAD), Carcinoma de próstata (PRAD), Carcinoma de vejiga (BLCA), Carcinoma de mama (BRCA), Carcinoma escamoso de pulmón (LUSC), Adenocarcinoma de estómago (STAD), Carcinoma de Útero (UCEC) y Cholangiocarcinoma (CHOL). Debido al tamaño de esta tabla, solo está disponible en la versión electrónica provista en el CD.

Tabla Suplementaria 8. Relaciones miARN-ARNm especificas enriquecidas la mayoría de los tumores.

Incluye el nombre del gen y la expresión de este (+ indica sobreexpresión y - represión), la ruta a la que pertenece el citado gen y el miARN con el cual interactúa y su expresión (+ indica sobreexpresión y - represión). Después se incluye en número de tumores donde ambos aparecen inversamente expresados y el valor de probabilidad obtenido por el test de Fisher. La última columna incluye aquellas interacciones publicadas por otros autores, NE significa No Encontrado.

Tabla Suplementaria 9. Valores de correlación en los pares conservados.

La siguiente tabla muestra los valores de correlación entre la expresión de los genes (logaritmo de RPKM) y la expresión de los microARNs (logaritmo de los RPKMs) corregidos por el efecto de la metilación y de la alteración en el número de copias. Un valor positivo implica una correlación positiva, mientras que un valor menor de cero implica una correlación negativa. El valor NA implica que o la expresión del gen o la del microARNs no pudo ser detectada en este tipo tumoral. Dado el tamaño de la tabla, se adjunta en el CD provisto.

Tabla Suplementaria 10. Valores de supervivencia de los pares conservados.

La tabla incluye para cada interacción encontrada (filas) en un determinado tipo tumoral (columna), los valores de riesgo relacionados con la supervivencia. Valores altos implica una mayor relación con un mal pronostico. Además se adjunta los valores de probabilidad ajustados por el método de FDR. Dado el tamaño de la tabla, se adjunta en el CD provisto.

Tabla Suplementaria 11. Valores de correlación para los pares exclusivos de pulmón.

La siguiente tabla muestra los valores de correlación entre la expresión de los genes (logaritmo de RPKM) y la expresión de los microARNs (logaritmo de los RPKMs) corregidos por el efecto de la metilación y de la alteración en el número de copias. Un valor positivo implica una correlación positiva, mientras que un valor menor de cero implica una correlación negativa. El valor NA implica que o la expresión del gen o la del microARNs no pudo ser detectada en este tipo tumoral. Dado el tamaño de la tabla, se adjunta en el CD provisto.

Tabla Suplementaria 12. Valores de supervivencia de los pares exclusivos de pulmón.
La tabla incluye para cada interacción exclusiva calculada (filas) en un determinado tipo tumoral (columna), los valores de riesgo relacionados con la supervivencia. Valores altos implica una mayor relación con un mal pronostico. Además se adjunta los valores de probabilidad ajustados por el método de FDR. Dado el tamaño de la tabla, se adjunta en el CD provisto.